



UNIVERSIDAD DE CANTABRIA

TRABAJO PARA OPTAR AL GRADO DE FÍSICA

Optimización de un método de detección de fuentes extragalácticas compactas en imágenes del Fondo Cósmico de Microondas

Optimization of a method for the detection of extragalactic
compact sources on images of the Cosmic Microwave
Background

José María Palencia Sainz

dirigido por

Diego HERRANZ MUÑOZ

10 de julio de 2020

Índice

1	Introducción	2
2	Datos utilizados: Imágenes del Fondo Cósmico del Microondas.	8
2.1	Almacenamiento y representación de datos del FCM	11
2.2	Datos Planck	12
3	Filtro utilizado: Filtro adaptado	13
4	Implementación del método de detección: Estructura y funcionamiento.	18
5	Presentación y análisis de resultados	25
5.1	Presentación de resultado y comparación con los PCCS2	25
5.1.1	Ajuste método RANSAC	35
5.1.2	Ajuste método ODR	38
5.2	Estudio de fuentes mono-catálogo	40
5.3	Número de cuentas en función del flujo	42
6	Tiempos de ejecución. Comparación con método en Fortran95.	46
7	Fases del proyecto	47
8	Hitos y conclusiones	48
A	Misión Planck	49

Palabras clave: Separación de componentes, Filtro adaptado, Fuentes compactas, Fondo Cósmico de Microondas, Cosmología, Astrofísica, Python.

Keywords: Component separation, Matched filter, Compact sources, Cosmic Microwave Background, Cosmology, Astrophysics, Python.

Optimización de un método de detección de fuentes extragalácticas compactas en imágenes del Fondo Cósmico de Microondas.

José María Palencia
Universidad de Cantabria.

10 de julio de 2020

Resumen

Un método de detección de fuentes compactas sobre imágenes de la Radiación del Fondo Cósmico de Microondas (RFCM), basado en el uso de filtros aplicados sobre la división de dichas imágenes en parches de emisión estadísticamente homogénea e isotropa, ha sido implementado en Python. Dicho método es una optimización de una versión más antigua, escrita en Fortran95. Este método ha sido utilizado sobre las imágenes a 30, 44 y 70 GHz de la RFCM publicadas en 2015 por la misión de la ESA Planck. Se ha realizado una comparación de los catálogos de fuentes compactas obtenidos con los publicados por Planck, obteniendo catálogos más extensos y con un alto grado de compatibilidad. Se ha realizado una comparación de las cuentas diferenciales en nuestro catálogo con los modelos teóricos actuales, comprobando la compatibilidad de nuestros datos con dichos modelos.

Abstract

A method for the detection of compact sources from images of the Cosmic Microwave Background radiation (CMB), based on the use of matched filters applied on patches of those images with isotropic and homogeneous radiation, has been implemented in Python. This method is an optimization of an older version, written in Fortran95. The method has been used on the 30, 44 and 70 GHz images of CMB published in 2015 by the ESA mission, Planck. A comparison between both the catalogues of compact sources obtained by us and those provided by Planck at each frequency has been carried out, obtaining larger catalogues in great concordance with those of Planck. A comparison between the differential counts of our catalogues and the actual theoretical models has been carried out with great concordance.

1 Introducción

La cosmología, ciencia encargada del estudio del Universo como un todo, su origen, composición y evolución, tiene como uno de sus pilares fundamentales el estudio del Fondo Cósmico de Microondas (FCM). Es en los últimos años, mediante el análisis de esta radiación primigenia por misiones como WMAP o Planck, ver [1] y [2], que se han realizado importantes avances en el campo.

El FCM es una radiación emitida en el Universo primigenio, cuando este tenía aproximadamente 400.000 años. Esta radiación es la primera que ha viajado libremente por el espacio, marcando un límite en nuestras observaciones. El hecho de no observar ninguna radiación previa al FCM lo convierte en nuestra fuente de información más antigua hasta la fecha. Debido a que su emisión fue tan temprana, las características del Universo joven influyeron completamente en sus propiedades. Estudiando las propiedades estadísticas del FCM como la polarización o intensidad de este, en distintas zonas del cielo, podemos obtener valiosa información para caracterizar y comprender, en mayor medida, al Universo que habitamos. Gracias al FCM sabemos que Universo es casi plano, tiene una edad aproximada de 14 mil millones de años, se expande de manera acelerada o, que la materia que nos forma solo constituye un 5 % del Universo. El problema

aquí recae en que dicha radiación, la cual ha recorrido el Universo por más de 13 mil millones de años, no nos llega tal y como se emitió. Esta radiación se encuentra en su camino con obstáculos como galaxias, polvo interestelar, nebulosas o la propia Vía Láctea. Sumado a esto, el movimiento de la Vía Láctea genera un efecto Doppler. La zona hacia la que se mueve la tierra la vemos desplazada hacia el azul y por tanto más caliente (energética), mientras que la radiación en la dirección opuesta del cielo está desplazada hacia el rojo y más fría. Por otra parte, la radiación del FCM, pese a haberse emitido a frecuencias muy energéticas, ha ido perdiendo energía con la expansión del Universo. Tal es este efecto, que hoy observamos esta radiación en las microondas. Debido a la existencia de fuentes en el Universo que emiten radiación de la misma frecuencia, es imposible observar únicamente al FCM. Estamos, irremediablemente, captando el resto de fuentes en nuestras imágenes. Estos efectos combinados afectan a las observaciones del FCM dificultando su estudio. Para conocer más sobre el FCM ver [3].

El trabajo de los cosmólogos en los últimos años ha sido en gran medida la corrección/eliminación de estas perturbaciones en las imágenes del FCM. Algunas como la corrección del efecto Doppler, producido por nuestro movimiento, son sencillas hoy en día. Otras en cambio no lo han sido tanto. Separar la contribución de otras fuentes de la radiación del FCM es un trabajo necesario. Dichas fuentes, de origen tanto galáctico como extragaláctico, se clasifican en dos grupos:

- Fuentes difusas. Estas fuentes no suelen presentar un borde claro que las delimite y ocupan una extensión angular sobre la esfera significativamente mayor que la resolución angular de nuestros telescopios. Este es el caso del propio FCM o el disco de nuestra propia galaxia.
- Fuentes compactas. En estas su señal se concentra mayoritariamente en una región compacta y menor que la resolución de nuestros telescopios. Se concentran en una región espacialmente pequeña. Un ejemplo claro es una galaxia en los experimentos modernos del FCM (Su tamaño angular suele ser menor que la resolución de los instrumentos de medida).

En este trabajo nos encargaremos de la detección de fuentes compactas.

Hasta ahora hemos tratado a las fuentes emisoras en microondas como ruido que "manchaba" nuestras imágenes del FCM. Lo que buscábamos era eliminarlas. Esto lo mencionábamos desde el punto de vista de la cosmología. Mirándolo desde otro ángulo, estas fuentes astrofísicas son importantes en sí para conocer más sobre los fenómenos que las caracterizan.

Durante décadas el análisis de las imágenes del cielo en microondas y ondas de radio ha contribuido enormemente a la astrofísica y cosmología.

- Desde el punto de vista de la astrofísica:

La radio astronomía extragaláctica experimentó un gran impulso tras el final de la II Guerra Mundial, cuando todas las antenas que se usaban para detectar aviones y misiles empezaron a usarse para observar el cielo.

La atmósfera no es igualmente transparente en todas las longitudes de onda. Es bastante transparente en el visible, pero muy poco en el ultravioleta, es muy opaca en el infrarrojo lejano, y en radio (que es un rango muy amplio) presenta distintos comportamientos. En frecuencias de unos pocos MHz la atmósfera es razonablemente transparente y se empezaron a detectar y estudiar las primeras radiogalaxias. En la zona de microondas, entre 20 y 300 GHz, la atmósfera absorbe especialmente bien la radiación, es muy opaca. Los primeros experimentos de FCM que miden en este rango buscaban evitar la absorción atmosférica (debida, principalmente, al vapor de agua). Por eso las antenas se construyeron en lo alto de montañas, en los polos, o incluso en el espacio. Teníamos desde entonces un buen conocimiento de las propiedades físicas de las radiogalaxias en frecuencias menores que unos pocos GHz, ver[4]. En la zona de microondas no teníamos una idea clara de como se comportaban las radiogalaxias.

Para conocer bien como funciona una galaxia, conviene poder observarla en la mayor cantidad de longitudes de onda. El "microwave gap" era un problema significativo, sabíamos como funcionaban las galaxias en el infrarrojo y en radio, pero no en el rango intermedio.

La zona de microondas separa dos regímenes muy distintos de fenómenos físicos que operan en las galaxias. En el infrarrojo (IR) la mayor parte de la radiación que vemos es térmica (de tipo cuerpo negro), debido al calor que tiene el

polvo que hay en las galaxias y a sus estrellas más pequeñas y frías. El IR nos da información acerca de la composición y evolución química de las galaxias, de cómo se forman estrellas en las mismas y de cómo son las zonas más frías. En radio no vemos emisión térmica, vemos emisión de sincrotrón: electrones libres que se mueven a grandes velocidades siguiendo los campos magnéticos de las galaxias. La radiación sincrotrón habla tanto de los mecanismos (normalmente violentos) que emitieron esos electrones (explosiones de supernovas, estrellas super calientes que ionizan el hidrógeno a su alrededor, púlsares, discos de acreción en torno a agujeros negros supermasivos y núcleos activos de galaxias, etc) como de los campos magnéticos en los que se mueven (flujos de plasma, movimientos cataclísmicos, ondas de choque, magnetosferas estelares y de núcleos galácticos, etc)

En la zona de las microondas no sabíamos que es lo que ocurría. Hay galaxias que emiten mucha radiación en radio pero poca en IR, en otras ocurre lo contrario. Otras galaxias emitían grandes cantidades de radiación en ambos regímenes. Las distinciones entre estas, si eran etapas diferentes en la evolución de las galaxias o, si dependía del medio en el que se encontraba la galaxia eran algunas de las preguntas sin respuesta que surgieron.

Completar nuestras observaciones en el dominio de las microondas, llenar el “microwave gap” era y es una necesidad en astronomía. Por esto la detección de fuentes compactas en imágenes del FCM, que se encuentra en estas frecuencias, ayuda enormemente al desarrollo de la astronomía.

El problema es que precisamente en frecuencias de microondas en donde las galaxias brillan menos. La curva típica de luz de una galaxia (spectral energy distribution, SED) suele presentar un máximo en el IR y otro en radio, con un mínimo justo entre medias: las microondas. Por eso observamos menos fuentes en los mapas de microondas. Es necesario desarrollar técnicas estadísticas de detección que nos permitan maximizar el número de galaxias detectadas.

- Desde el punto de vista de la cosmología:

Los radiocatálogos de mediados y finales del siglo XX proporcionaban buenas selecciones de candidatos entre los que encontrar objetos a alto z (más lejanos). Esto es debido a que algunos cuásares y otros AGNs tienen una potente emisión de radio. Un cuásar parece una estrella (quasi-stellar). En el óptico, salvo que observes las líneas de emisión de su espectro, es complicado distinguirlo entre otros millones de estrellas (es fácil que pase inadvertido). Sin embargo, en radio destaca frente a los demás objetos con pinta de estrella. Por eso los radiocatálogos eran (y siguen siendo) una buena herramienta de selección de candidatos, que luego se exploran más detenidamente en el resto de bandas.

La era de mayor actividad de cuásares fue hace unos diez mil millones de años ($z \propto 2$). Usando radiocatálogos era (y sigue siendo) relativamente sencillo poder localizar objetos de cuando el universo tenía la cuarta parte de su edad actual.

Las mayor parte de las fuente que observábamos por los telescopios ópticos estaban a distancias menores. La población de galaxias descubiertas en el rango visible daba un tamaño aparente del Universo menor del que tenía. Con las radio antenas detectábamos objetos mucho más lejanos, galaxias a alto corrimiento al rojo z .

El corrimiento al rojo se mide gracias a la espectrografía de estas fuentes. Hallando líneas de emisión las cuales conocemos el laboratorio (Ejemplo: línea K_α del Fe) y comparándolas se tiene que:

$$1 + z = \frac{\nu_{laboratorio}}{\nu_{observada}}, \quad (1)$$

donde ν representa la frecuencia de la línea.

La ley de Hubble-Lemaître puede expresarse tal que:

$$z = \frac{H_0}{c} D, \quad (2)$$

donde H_0 es la constante de Hubble, c es la velocidad de la luz y D la distancia propia en megapársec. Podemos entonces relacionar el corrimiento al rojo z con la distancia a la galaxia. El z que aquí mencionamos es el cosmológico. Existen otros corrimientos al rojo: uno gravitacional (la luz pierde energía/frecuencia al escapar de un campo gravitacional), y otro Doppler (el movimiento propio de la galaxia lo genera). El cosmológico se debe a la expansión del Universo. Esta expansión genera una pérdida de energía en la radiación (esto es lo que le ocurrió a la radiación del FCM). Para poder utilizar esta ley necesitamos que el z cosmológico domine sobre el resto. Esto ocurre para galaxias muy cercanas donde

la velocidad de recesión debida a la expansión del Universo domina. Las galaxias más lejanas presentan un corrimiento al rojo cosmológico mayor, por esto las galaxias detectadas por nuestras radioantenas se encuentran a distancias muy lejanas (z cosmológico domina). Gracias a las radiogalaxias detectadas se pudo comprender mejor la escala del Universo.

Por otra parte, el descubrimiento del FCM por Penzias y Wilson [5] del FCM en 1965 acabó por demostrar el Big Bang y desmontó la teoría estacionaria. Durante muchos años hubo, entre los astrónomos, un debate sobre el carácter del Universo. Unos mantenían que este no tuvo origen, si no que siempre existió de la misma manera como lo conocemos hoy. Otros en cambio mantenían que el Universo tuvo un origen y que este evoluciona. Algunos sugerían que este se estaba expandiendo. Retrocediendo en el tiempo este se contraería hasta reducirse a una singularidad en el inicio. A este principio se le dio en llamativo nombre de *Big Bang*. El descubrimiento del FCM contribuyó a zanjar este debate.

El estudio de las propiedades estadísticas del FCM ha permitido medir, entre otras cosas: la curvatura del Universo, su edad o su composición. El Universo esta compuesto por un 5 % de materia bariónica (la que nos forma a nosotros, galaxias y estrellas), un 27 % de materia oscura (no sabemos que es, solo que su única interacción conocida es la gravitatoria), y un 68 % de una energía oscura causante de la expansión acelerada del Universo. Los parámetros cosmológicos principales son: H_0 , Ω_Λ , Ω_m , Ω_r , Ω_k . Estos parámetros principales informan sobre las propiedades anteriormente mencionadas (edad, composición, curvatura...). Sin embargo existen otros parámetros cosmológicos “secundarios” como el número de familias de neutrinos, la profundidad óptica de reionización (que nos habla de cuándo se encendieron las primeras estrellas del Universo) o el índice espectral del espectro de potencias primordial de las fluctuaciones de densidad de energía/materia pueden ser deducidas del estudio de las fluctuaciones de temperatura y polarización del FCM. La herramienta principal para obtener estos parámetros es el estudio del espectro angular de potencias de las fluctuaciones de temperatura del FCM. No indagaremos en esto debido a que nuestro trabajo se enfoca en la detección de fuentes extragalácticas.

Junto a estos datos podemos crear modelos sobre la futura expansión del Universo o interpolar a momentos intermedios entre la emisión del FCM y ahora. Mediante el estudio de imágenes a mayor resolución y mejor filtradas podemos reducir los errores de medida mejorando la precisión de los resultados. Por otra parte hay teorías como el modelo inflacionario que predicen pequeñas variaciones en el FCM. Por ejemplo, el modelo estacionario predice un fondo cósmico de ondas gravitacionales anteriores al FCM, producidas cuando el Universo tenía solo una fracción de segundo de vida. Para comprobar o descartar estos modelos necesitamos de una enorme precisión. Imágenes más precisas del FCM pueden ser la puerta a nueva física y cambios en nuestro concepto del Universo primordial.

Las poblaciones de fuentes de radio incluyen galaxias a alto z (galaxias muy lejanas aumentadas por efecto de lente gravitacional), radio AGNs, galaxias con formación de estrellas y proto-cúmulos de galaxias (comienzo en la agrupación de cientos o miles de galaxias). Estudiar este tipo de fuentes nos permite analizar y comprender como se forman las galaxias, sus primeras etapas, como se forman los cúmulos de galaxias o, cómo son el interior de las galaxias con núcleos activos (AGN). Existe una gran cantidad de información muy valiosa en el rango de microondas que queremos aprovechar y necesitamos para entender mejor estas fuentes. Por otra parte aumentar la estadística de las radiogalaxias permite cerrar el “microwave gap”. Un análisis detallado de las poblaciones de fuentes de radio-microondas y su importancia viene dado en [6].

Desde el punto de vista de los radio astrónomos es el FCM de microondas el que tratamos como ruido. El FCM contamina nuestras observaciones de las mencionadas fuentes. Ambas señales entremezcladas (FCM + radio fuentes) son importantes y, sería lo ideal separarlas conservando ambas contribuciones lo más fieles a la realidad posible.

Para la separación de fuentes difusas lo más utilizado estos últimos años han sido las técnicas Blind Source Separation (BSS), separación de fuente a ciegas. Este conjunto de métodos separan fuentes mezcladas entre sí, sin conocer o conociendo muy poco acerca de las señales o como se han mezclado. Se puede ver en [7] una comparación de esta con otras técnicas en la misión Planck.

Las fuentes compactas pertenecen a tres grupos: galaxias muy lejanas (radio, galaxias, cuántares, blazars, galaxias ordinarias, proto-galaxias, galaxias de polvo con gran formación de estrellas y muchas otras); clusters de galaxias (alteran la radiación del FCM por efecto Compton inverso); y objetos galácticos compactos (núcleos fríos, remanentes de supernova, y concentraciones de polvo caliente). Aparecen distribuidos de manera homogénea por el cielo, salvo el tercer tipo que se acumula en el plano galáctico. Salvo algunos clusters de galaxias todos tienen tamaños angulares menores que la resolución de los experimentos actuales de FCM (estos clusters de galaxias podrían ser fuentes difusas). Debido a la gran diversidad de fuentes compactas muchas de las técnicas de separación de objetos fallan.

La detección de fuentes compactas en astronomía es un problema antiguo para el que existen varias soluciones. Para encontrar objetos compactos en imágenes se han ido creando algoritmos de detección con el paso de los años. Con cada nueva ventana electromagnética se han creado o actualizado estos algoritmos. Por ejemplo el algoritmo CLEAN [8] se ha utilizado en la radioastronomía por 35 años. Otro como SEXTRACTOR [9] ha sido útil en el visible, IR y astronomía de rayos X. Ninguna de estos métodos (salvo quizás SEXTRACTOR) ha sido de gran utilidad en la detección de fuentes compactas a las frecuencias del FCM.

En las imágenes del FCM no podemos diferenciar a priori las fuentes compactas del FCM y otro tipo de ruidos. Todo el ruido muestra una correlación espacial (ruido de color) salvo por el ruido del detector. El método CLEAN encuentra la posición e intensidad de una fuente situándose en el punto de máxima intensidad y comparándolo a un perfil, lo elimina de los datos y repite el procedimiento hasta cumplir un criterio de convergencia. CLEAN asume que el cielo se compone de unas pocas fuentes y un fondo casi nulo. Esto no ocurre en el FCM donde las fluctuaciones del fondo pueden ser similares en intensidad a las fuentes y dominan en número. CLEAN está diseñado para el FCM. SEXTRACTOR aproxima la emisión de fondo a un polinomio de orden bajo que se sustrae de la imagen. Las fuentes compactas se obtienen al encontrar grupos de píxeles por encima de un umbral de intensidad. La emisión del FCM varía en una escala angular del orden de los minutos de arco, similar a la resolución de nuestros experimentos. Por esto no podemos aproximar el fondo a un polinomio de orden bajo. SEXTRACTOR no está optimizado para este tipo de problema. Es, por tanto, necesario desarrollar métodos específicamente diseñados para abordar las condiciones particulares que aparecen en las imágenes del FCM.

El problema que tenemos es el siguiente: tenemos un número desconocido de fuentes compactas débiles, esparcido homogéneamente por todo el cielo, y con tamaños angulares comparables a la variación del FCM que cubre todo el cielo. Este ruido (FCM + instrumental + componentes galácticos) es prácticamente gaussiano, salvo por la componente galáctica que esta distribuida de una manera fuertemente no gaussiana. En este contexto todos los métodos antiguos usados en el óptico, IR, o rayos X no funcionan de forma óptima. Necesitamos un nuevo filtro.

Por motivos que se explicarán más adelante, al ser la fuentes de un tamaño angular mucho menor que la resolución del instrumento, sabemos que las fuentes ser observan en nuestros datos como una campana gaussiana (geometría prácticamente idéntica a la respuesta angular de nuestra antena). Esto hace que incluso si no sabemos el perfil de la fuente, una vez medido sabemos la forma que va a tomar. En cuanto al ruido no lo conocemos tanto como nos gustaría. Sabemos que salvo en la zona en que el ruido dominante proviene de la galaxia, la aproximación de que la PDF, Probability Density Function (función de densidad de probabilidad), es gaussiana es acertada. La teoría de la inflación predice que las fluctuaciones de temperatura del FCM son gaussianas o casi gaussianas (hay teorías de inflación un poco menos estándar en las que se predice un pequeño grado de no-gaussianidad). El ruido instrumental de los radiómetros y bolómetros que se utilizan en experimentos de FCM se debe en parte a la electrónica del detector que tiende a crear ruido gaussiano. Esperamos que la suma de FCM y ruido sea razonablemente gaussiana. Además, cuando empiezo a sumar señales aleatorias (FCM + ruido + sincrotrón galáctico + emisión free-free + spinning dust + thermal dust + ...) empieza a operar el Teorema Central del Límite. Este nos dice que la suma de procesos aleatorios independientes acaba llevando a una PDF total que tiende rápidamente a la gaussianidad.

Conocemos el perfil que tomarán las fuentes compactas en nuestras imágenes. Sabemos que lejos de la galaxia la PDF del ruido es prácticamente gaussiana y en la galaxia fuertemente no gaussiana. Queremos, ya que sabemos el perfil de la fuente, un filtro que respete en las imágenes los datos que sigan el perfil de la fuente, manteniendo el valor del píxel central (intensidad de la fuente). Por otro lado queremos que el ruido, que no sigue el perfil de la fuente, se elimine o se reduzca (intensidad). Para poder obtener catálogos con un gran grado de fiabilidad (pocas falsas detecciones) y completos (que detecten la mayor cantidad de fuente), necesitamos un criterio de detección. El método de detección que mejor se ajusta a nuestros requisitos es un filtro adaptado complementado de un umbral de detección. Más tarde se desarrollará el funcionamiento de este. Para conocer más sobre la detección de fuentes compactas, ver [10].

La técnica de filtros adaptados es aquella en la que conocemos el perfil de la fuente. Aplicamos sobre la imagen este filtro: Todo aquello que cumpla el perfil mantiene su valor, lo que no tenga el perfil ve reducido su amplitud (intensidad). Este método maximiza la relación entre el valor de la señal, que cumple el perfil, con el fondo. Ver imágenes 1a y 1b. Esto por si solo no nos indica si hemos detectado una fuente compacta o no. Queda complementarlo con un valor de corte que estime la relación entre el valor de un pico filtrado sobre el ruido, un umbral para considerarlo una detección. Más sobre el funcionamiento del filtro, sus problemas y ventajas será discutido más adelante.

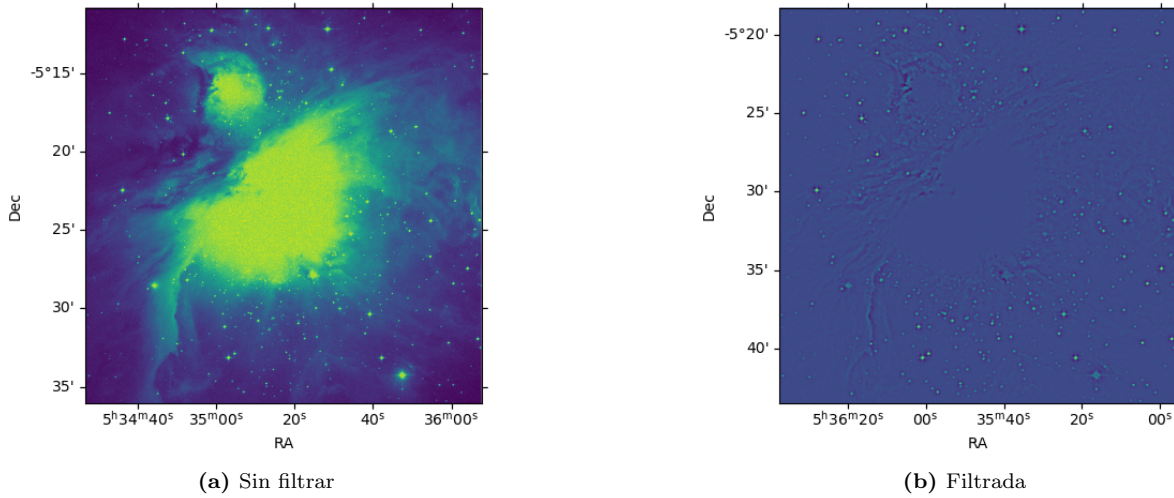


Figura 1: Imagen con fuentes compactas a la cual se le ha aplicado la técnica de filtros adaptados.

Se puede deducir de todo esto que el uso de técnicas como la de filtros adaptados es doblemente útil. En el campo de la cosmología, para poder ubicar y posteriormente eliminar las fuentes compactas de las imágenes del FCM. Para el campo de la astrofísica, ya que se pueden obtener catálogos de los distintos tipos de fuentes compactas que aparecen en las imágenes, las cuales, tienen su propio interés. Un análisis sobre el mutuo beneficio de ambos campos utilizando estas técnicas viene desarrollado en [11].

Este trabajo tiene como motivación la implementación de un código en Python que, utilizando la técnica de filtros adaptados, genere a partir de imágenes del FCM catálogos de fuentes compactas. Dichos catálogos incluirán datos como con sus coordenadas o estimaciones de flujo. El trabajo se basa en un método similar implementado en Fortran95 por colaboradores de la misión Planck, entre ellos el director de este trabajo, con el cuál se generaron los catálogos de fuentes compactas publicados en 2015 por la colaboración Planck.

La necesidad de este trabajo es entre otras cosas, es el cambio a un lenguaje de programación en auge como es Python. Python presenta varias ventajas sobre Fortran95. Es más “user friendly”, lo que lo hace más atractivo a iniciados en la programación. Es libre y gratuito, mientras que muchos de los compiladores de Fortran son de pago, no todos. Es un lenguaje orientado a objetos, presentando una mayor cantidad de recursos. Dispone de multitud de librerías científicas convenientes para el análisis y representación de datos. Finalmente, y lo más importante, Python se está convirtiendo en uno de los lenguajes de programación científica más utilizados. Buscamos al cambiar a Python implementar nuestro código para que pueda ser utilizado por cualquier persona en cualquier ordenador. Queremos subir este código a un repositorio para su libre uso de cara a futuros experimentos. Buscamos por otra parte, generar catálogos de fuentes compactas más extensos que los publicados por Planck, para su uso por astrónomos y cosmólogos.

Uno de los puntos novedosos de este trabajo es que, independientemente de los datos aportados, está programado de manera que dado cualquier mapa del FCM, tomado a cualquier frecuencia y, con cualquier antenna con su resolución característica, puede obtener los catálogos de fuentes compactas. Esto hace que además de ser útil para trabajar sobre los datos de Planck, pueda hacerlo también sobre futuros datos de misiones como QUIJOTE, ver [12]. La necesidad de este trabajo es, en conclusión, la de optimizar un método ya existente y probado de detección de fuentes compactas en imágenes del FCM, su preparación para posibles usos futuros en misiones que requieran de la limpieza de fuentes compactas en sus imágenes, así como para la creación de catálogos más extensos que permitan hacer un mejor análisis estadístico de las propiedades de este tipo de fuentes.

La motivación personal de este trabajo es la de obtener una formación en competencias básicas para la investigación en ciencia. Algún ejemplo de esto puede ser el análisis de datos, el diseño de métodos para resolver problemas o, la optimización de las herramientas utilizadas. Por otra parte se pretende también conseguir una familiarización con las herramientas y métodos, así como bases de datos principales, en el campo de trabajo de la astrofísica y cosmología.

También se pretende conocer más sobre los experimentos pasados, presentes y futuros de dichos campos. Todo esto con el objetivo de obtener la formación básica necesaria para empezar una vida profesional dedicada a la investigación.

En las páginas contiguas se explicará en detalle el tipo de datos utilizados, una descripción más completa de la técnica utilizada (filtros adaptados), una explicación detallada del funcionamiento del código creado y de las principales librerías de la comunidad utilizadas. Seguirá la presentación de los resultados obtenidos, así como un exhaustivo análisis de estos.

2 Datos utilizados: Imágenes del Fondo Cósmico del Microondas.

Como ya se ha explicado antes, los datos sobre los que trabajamos son imágenes del FCM. El fondo cósmico de microondas es una radiación de microondas proveniente de todas las direcciones del espacio. Esta radiación proviene de los primeros instantes del Universo. La teoría más aceptada hasta la fecha es que tras el Big Bang el Universo, mucho más pequeño que en la actualidad, se encontraba dominado por la radiación. El Universo en estos momentos es un plasma. Los fotones tienen una gran energía, tal que en el momento en que un protón y un electrón se acoplan formando hidrógeno se absorbe uno de estos fotones, esto acaba rompiendo el enlace. La materia y radiación están acopladas, no se forman especies atómicas y la radiación no viaja libremente, ya que es rápidamente absorbida. Debido a esto la radiación emitida en aquellos momentos no nos llega. Pese a estar dominado por la radiación este es un universo “oscuro” para nosotros.

Con la expansión del Universo, la radiación se enfría. Llegado un punto, aproximadamente 380.000 años después del Big Bang, la radiación no tiene energía suficiente para romper los enlaces atómicos. La radiación se desacopla de la materia y se forman los primeros átomos. Este hecho se conoce como reionización. Desde cada punto del espacio la radiación se emite de manera isotrópica hacia todas las direcciones del espacio. En cada punto del espacio observamos esta radiación proveniente de todas direcciones, detrás de la cual no nos llega ninguna radiación, por ello el nombre Fondo Cósmico. Desde su emisión hasta hoy, más de 13 mil millones de años después, el Universo se ha expandido. La radiación emitida en aquel entonces perdió energía debido a la propia expansión del Universo. La única manera en que una radiación electromagnética pierda energía es en forma de frecuencia, por ello hoy el FCM cae precisamente en el límite entre las microondas y las ondas de radio y por esto la tercera componente del nombre. Este se comporta como un cuerpo negro perfecto a una temperatura de $2,72548 \pm 0,00057$ K [13], cuyo pico de radiación cae entorno a una frecuencia de 160,23 GHz. Pese a que la cantidad física que medimos es la de densidad de flujo espectral (Energía por unidad superficie y unidad de frecuencia), normalmente medida en janskys (Jy), unidad no perteneciente al Sistema Internacional de unidades (SI), los mapas del FCM se suelen expresar en temperatura, en kelvin (K). Esta temperatura es una equivalente, es la temperatura que tendría un cuerpo negro cuyo máximo de densidad espectral coincide con el que medimos.

El FCM es isotrópico y homogéneo. Un observador en cualquier lugar del Universo observa las mismas propiedades estadísticas en cada zona del espacio (Homogéneo). El Universo presenta las mismas propiedades en todas las direcciones (Isótropo). Pese a esto presenta pequeñas variaciones locales de temperatura conocidas como las anisotropías del fondo cósmico de microondas, figura 2. Las anisotropías son del orden de unas pocas decenas a centenas de micro-kelvins (μK). Hay dos tipos de anisotropías, las primarias y las secundarias. Las secundarias se debe a las interacciones de la radiación con los distintos cuerpos en el Universo tras su emisión. Estas anisotropías pueden ser contaminaciones de los datos que buscamos corregir, como la reionización que quita o aporta fotones al medio, o alteraciones como la influencia de la materia creando efectos de lente gravitacional o el efecto Salch-Wolfe integrado (ISW) que añade un corrimiento al rojo gravitatorio. La primaria procede de las propiedades del Universo joven que influyeron en la emisión del FCM. A partir del estudio de las anisotropías primarias podemos obtener información del Universo en los instantes de la reionización e incluso anteriores. Con esa información podemos obtener datos sobre el Universo actual. Las anisotropías primarias se dividen en dos grupos: oscilaciones acústicas y dispersión amortiguada. La primera de estas dos se debe a la lucha entre la atracción gravitatoria de los bariones (partículas subatómicas del plasma primigenio) y la presión de radiación. Esto creó sobredensidades y subdensidades de materia en distintos puntos del espacio. La dispersión amortiguada se debe a la difusión de fotones de zonas más calientes a más frías, intentado alcanzar el equilibrio térmico. Estas sobredensidades acabaron germinando en las primeras galaxias mientras que las subdensidades generaron vacíos. Todo esto terminó causando la estructura a gran escala del Universo actual. La magnitud de estas anisotropías permiten obtener información del Universo, por ejemplo, de las oscilaciones acústicas hemos obtenido la curvatura del Universo,

siendo este casi plano. El FCM también está polarizado debido a diversos mecanismos, como la dispersión Thompson con las partículas del plasma.

Tanto Planck como otros experimentos utilizan antenas en tierra o en satélites orbitando a la Tierra que barren el espacio en diferentes frecuencias, obteniendo mapas esféricos del FCM (El FCM llega a la tierra desde todas las direcciones). Miden su intensidad en K y su polarización. Hay dos tipos fundamentales de detectores que se usan en microondas: radiómetros y bolómetros. Los bolómetros miden (indirectamente) temperatura, pero los radiómetros (lo que se usa por debajo de 100 GHz. La tecnología de bolómetros de baja frecuencia está poco desarrollada) son antenas de radio y no miden temperatura, sino flujo espectral (potencia electromagnética por unidad de frecuencia/longitud de onda, W/Hz o W/m). Lo que se hace es hablar de "temperatura de brillo" (temperatura de un cuerpo negro que emitiría esa misma cantidad de radiación). En el campo del FCM, también usamos otra temperatura relacionada "temperatura termodinámica". Mediante la ley de Rayleigh-Jeans (aproximación de la ley de Planck del cuerpo negro en bajas frecuencias, dominio en el que nos movemos) podemos obtener una equivalencia entre temperatura y densidad de flujo:

$$I_\nu = \frac{2\nu^2 kT}{c^2}, \quad (3)$$

donde I_ν es la densidad de flujo espectral para la frecuencia ν , k es la constante de Boltzmann, T la temperatura y c la velocidad de la luz. Para la detección de fuentes compactas sólo es necesaria la temperatura del FCM. La polarización es igualmente importante para los análisis del FCM, pero no influye en nuestro trabajo.

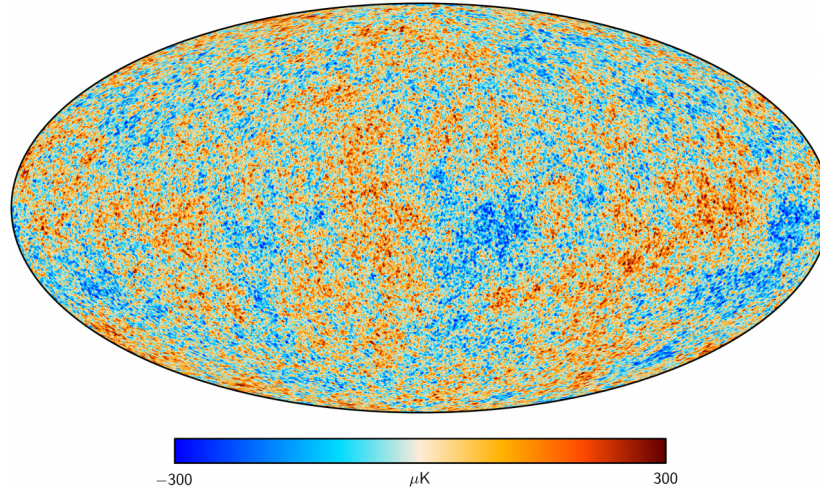


Figura 2: Mapa de las anisotropías del FCM obtenido de las imágenes del satélite Planck.

Todas la fuentes que emitan radiación en los rangos de frecuencias en los cuales barremos el cielo van a ser, inevitablemente, captados por los experimentos.

Una señal puede ser descrita matemáticamente como una amplitud multiplicada por un perfil que la caracterice:

$$s(\vec{x}) = A\tau(\vec{x}), \quad (4)$$

donde, s es el valor de la señal en el punto del espacio \vec{x} , A es la amplitud (intensidad de la señal) y τ es el perfil característico (forma de la señal). Si el perfil fuera simétrico tendríamos entonces:

$$s(\vec{x}) = A\tau(r), \quad (5)$$

donde r es la distancia al centro de la fuente. Podemos describir ahora los datos que medimos como la suma de nuestra señal $s(\vec{x})$ más un ruido $n(\vec{x})$ de forma que:

$$d(\vec{x}) = s(\vec{x}) + n(\vec{x}), \quad (6)$$

Es necesario hablar ahora de la respuesta angular de la antena o beam, b . Esta función caracteriza al instrumento que mide la señal. Inevitablemente la óptica de nuestros instrumentos que recogen la señal, dejan implícito en los datos

registrados una huella característica de sus propiedades físicas, como su forma, o tamaño que modifican la señal. La respuesta angular de la antena es perfectamente conocida. Sabemos que los datos que registramos no son más que el producto de convolución de estos con la respuesta angular de la antena:

$$\tilde{d}(y) = \int d(x)b(x-y)dx. \quad (7)$$

No observamos la señal como realmente es. Observamos una señal modificada por el beam de la antena, b . Tenemos que tener esto en cuenta. Si vamos a usar un filtro adaptado necesitamos indicar el perfil de las fuentes convolucionadas con el beam.

Debido al límite de difracción, límite físico que no podemos eliminar, ni siquiera en el mejor de los casos en el que contáramos con un instrumento perfecto, vamos a evitar este fenómeno. Tenemos que la imagen de una señal puntual se verá como una mancha. Esto impone en cierta medida un radio entorno a una fuente puntual. Si tenemos otra fuente dentro de este radio no la vamos a discernir, no seremos capaces de resolverlas. Podemos reducir este límite de dos formas: aumentando el tamaño de nuestro instrumento o aumentando la frecuencia a la que medimos. Esto genera un problema al tratar con imágenes del FCM. Tratar con frecuencias tan bajas aumenta el límite de resolución. Esto hace que las imágenes tengan menor resolución que imágenes tomadas a frecuencias mayores y con instrumentos similares en tamaño. Esto es un problema en el campo de las microondas ya que usando los instrumentos actuales, la FWHM (anchura a media altura de nuestro instrumento), algo así como un límite de resolución, por debajo del cual no podemos discernir varias fuentes, ver 3. En experimentos como WMAP la FWHM es del tamaño de las fluctuaciones del FCM, esto dificulta la detección de fuentes compactas en estas imágenes.

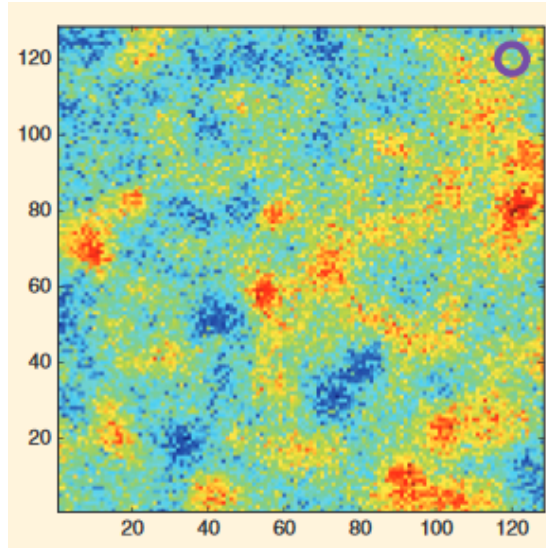


Figura 3: Porción cuadrada del cielo de 14.63×14.6 grados, centrada en coordenadas galácticas, $b = 60^\circ$ y $l = 100^\circ$, observadas por el satélite WMAP a 23 GHz. En la esquina superior derecha hay superimpuesto una circunferencia cuyo diámetro es la FWHM del aparato para esa frecuencia. Fuente: [10]

Cuando el perfil de la fuente es mucho más pequeño que la respuesta angular de la antena ($\tau \ll b$), es como si τ fuera una delta de Dirac:

$$\tilde{s}(y) = \int s(x)b(x-y)dx \approx \int \delta(x)b(x-y)dx. \quad (8)$$

La convolución de una delta de Dirac con una función, en este caso b , es igual a la misma función. Obtenemos de la ecuación 8 que

$$\tilde{s}(y) \approx b(y). \quad (9)$$

En el caso de las fuentes extragalácticas compactas (cuyo tamaño angular suele ser de unos pocos segundos de arco) observadas con los beams típicos de microondas (del orden del minuto de arco o superior) se cumple este criterio. El

perfil que observamos es idéntico al beam del instrumento. Podemos utilizar el filtro adaptado conociendo el beam pero sin conocer τ . Aun si la fuente tiene un perfil específico τ , no nos importa para la detección si se cumple que $\tau \ll b$. En este caso todas las fuentes compactas son iguales, salvo por su posición y amplitud, pudiendo utilizar el mismo diseño de filtro para todas ellas (las fuentes tienen el mismo perfil, solo cambia el ruido de una zona a otra). Esto simplifica enormemente nuestra tarea. Las fuentes difusas al no estar delimitadas en un espacio pequeño no cumplen ese criterio, cada una convolucionada con el beam mostrando una forma distinta.

2.1 Almacenamiento y representación de datos del FCM

Una vez obtenidas las imágenes esféricas del cielo poco energético, en las microondas, queda decidir cómo almacenar ese gran conjunto de datos. La solución utilizada por la misión Planck, cuyos datos hemos utilizado, fue la de utilizar el método HEALPix (Hierarchical Equal Area isoLatitude Pixelation), ver [14]. Este método divide la esfera en píxeles irregulares que cubren toda la esfera.

HEALPix nació de la necesidad de la misión WMAP y posteriormente Planck, de ordenar los datos tomados del FCM a diferentes frecuencias. Requerían de una estructura matemática que permitiera discretizar funciones en la esfera con la suficientemente alta resolución y, que permitiera tomar rápidos y precisos análisis estadísticos sobre los enormes conjuntos de datos obtenidos. Debido a estas cualidades, este método pronto pasó a ser utilizado en áreas totalmente distintas al análisis de mapas del FCM, para el cual fue originado, como topografía terrestre. HEALPix posee las siguientes propiedades¹:

1. La esfera está teselada jerárquicamente en cuadriláteros curvilíneos. La resolución mínima es de 12 píxeles, ver figura 4
2. El área de todos los píxeles es la misma.
3. Los píxeles se distribuyen en líneas de latitud constante. Esto acelera el cálculo de los armónicos esféricos, tremendamente importante en el análisis de datos en el FCM.

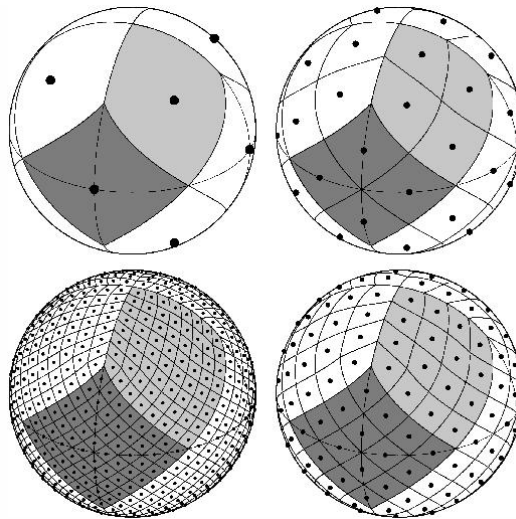


Figura 4: Esfera teselada con resoluciones de 12, 48, 192, y 768 píxeles siguiendo el sentido de las agujas del reloj.

HEALPix está implementado en diferentes lenguajes de programación como C, C++, Fortran90, IDL, Java o Python. Cuenta con rutinas que permiten manipular los píxeles en que divide la esfera, aplicar máscaras, visualizar los

¹<https://healpix.sourceforge.io/>

datos o, aplicar transformadas con armónicos esféricos entre otras. La implementación de este método en Python se conoce como Healpy² [15] y es uno de los paquetes principales utilizados en este trabajo.

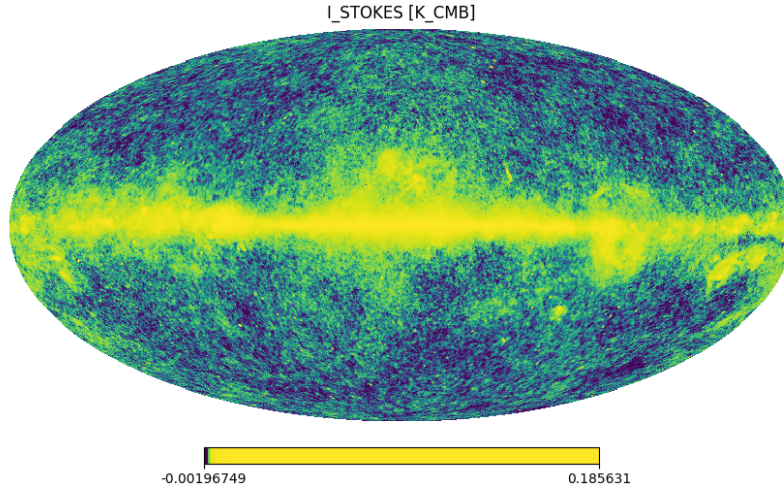


Figura 5: Representación de los datos obtenidos por Planck de la intensidad del FCM, representada por Healpy en la proyección Mollweide.

Quizás el aspecto más importante para este trabajo es que Healpy nos permite proyectar en parches cuadrados regiones de un área dada del cielo. De esta forma podemos obtener a partir de imágenes esféricas del FCM un conjunto de imágenes cuadradas sobre las cuales aplicar nuestro filtro. Podríamos aplicar nuestro filtro sobre todo el conjunto de mapas sin parchear pero, por motivos que se comentarán más adelante, es contraproducente.

Otro punto a tener en cuenta es en que formato se almacenan estos mapas con la estructura de HEALPix. La respuesta es en el formato FITS, Flexible Image Transport System. El formato FITS es un formato de almacenamiento de imágenes como JPG, PNG, GIF u otros más utilizados para guardar imágenes. Este es, en la actualidad, el formato estándar en astronomía para guardar datos o imágenes. Permite guardar en un mismo archivo varias imágenes, tablas complementarias o, incluso descripciones sobre los datos. Además son fáciles de manipular. Los archivos FITS suelen venir acompañados de una extensión llamada cabecero. El cabecero da información sobre los datos almacenados: cuando se tomaron, donde y en que instrumento se tomaron, las unidades de los datos de las imágenes y otros datos descriptivos más.

Los datos utilizados, los cuales vienen en formato fits, son leídos gracias a un paquete diseñado por la comunidad para Python, Astropy. Astropy, permite leer estos ficheros fits y almacenarlos en la memoria del ordenador. Cuenta además con utilidades como asignación de unidades (permite un uso sin preocuparse de realizar las transformaciones de unidades), cambios de coordenadas de un sistema a otro, transformación de coordenada a píxel en una imagen y viceversa, cálculo de distancias angulares o búsqueda de objetos en un radio dado. Astropy es, sin lugar a dudas, el paquete fundamental para astronomía en Python.

2.2 Datos Planck

Entendiendo mejor ahora el tipo de datos con el que hemos trabajado, sus propiedades, como se almacenan y, las librerías con las que utilizarlos, queda mencionar que datos hemos utilizado en este trabajo.

De los dos instrumentos de Planck, el LFI y HFI, instrumentos de baja y alta frecuencia respectivamente, nosotros, por comodidad, hemos utilizado los del LFI. Los mapas del LFI son mapas más pequeños permitiendo hacer pruebas más rápidas y sencillas. El objetivo es realizarlo de manera que dado un mapa del FCM tomado con el HFI o con otro instrumento funcione de igual manera.

²<https://healpy.readthedocs.io/en/latest/>

Por otro lado los catálogos de fuentes compactas que hemos utilizado para comparar los obtenidos por nosotros son los PCCS2, Second Planck Catalog of Compact Sources. Los PCCS2 son los últimos catálogos de fuentes compactas obtenidos por la misión Planck, sobre los que está basado este trabajo, que intentamos mejorar. Tanto los mapas usados como los catálogos que hemos utilizado para comparar los nuestros, se pueden obtener sin ningún requisito del Planck Legacy Archive³, donde todos los datos obtenidos por el experimento están publicados para el uso de cualquiera que los necesite.

3 Filtro utilizado: Filtro adaptado

La idea a la hora de obtener nuestros catálogos de fuentes compactas es simple. Tenemos una imagen del FCM de la que sabemos: como ha sido obtenida, que características tiene el instrumento con el que se ha realizado la medida y, como son las fuentes que esperamos obtener. Queda utilizar una herramienta la cual usando estos datos nos permita reducir el valor del ruido, en este caso el propio FCM, y preserve la amplitud de las fuentes en el centro (Preserva la densidad de flujo de la fuente).

Las fuentes compactas que medimos junto al FCM aparecen en nuestras imágenes como sobredensidades, zonas del cielo más brillantes. Podríamos pensar, de manera ingenua, que los píxeles en una zona dada del cielo, con una intensidad igual a un número de veces valor medio del ruido, tendría que ser una fuente compacta. En algunos casos tendríamos razón, pero esto no nos sería útil. Varias fuentes compactas podrían ser lo suficientemente débiles como para que no pasaran nuestro umbral de detección. Peor aún, alguna fluctuación del ruido podría superar este umbral pasando por una fuente compacta. En definitiva, no solo nos perderíamos fuentes, si no que tomaríamos por tales a fluctuaciones del fondo que no son más que ruido. Es necesario algo más que imponer un valor umbral para la detección. Necesitamos combinarlo con una herramienta que nos permita dejar intactas a nuestras fuentes y que reduzca el valor del ruido.

Sabemos que las fuentes compactas que medimos adquieren, una vez pasado por nuestros instrumentos, un perfil idéntico al beam. El beam es más o menos gaussiano, cuya anchura viene caracterizada por la frecuencia de la radiación y el tamaño de la antena. Usando ahora un filtro adaptado a este perfil, podemos aplanar el ruido maximizando el cociente entre la intensidad de la fuente y la dispersión del ruido (Signal to Noise Ratio SNR). El SNR es una medida de como destaca la fuente en la imagen.

El filtro adaptado funciona sólo en casos en los que tenemos una o varias señales en un conjunto de datos dominado por el ruido. En nuestro caso el FCM domina sobre las fuentes compactas.

El filtro adaptado es la correlación cruzada de nuestros datos con el perfil que conocemos que tendrán las fuentes compactas. La correlación cruzada es clave para la detección de patrones en señales mezcladas y es ampliamente utilizada y reconocida en múltiples campos, no solo en la astronomía y cosmología. Se define la correlación cruzada de una función $b(\vec{x})$ con otra función $d(\vec{x})$ como:

$$(b * d)(\vec{x}) = \int_{-\infty}^{\infty} b(\vec{y})^* d(\vec{x} + \vec{y}) d\vec{y}, \quad (10)$$

donde b sería nuestro perfil y d nuestros datos. b^* denota el complejo conjugado, sería idéntico al producto de convolución de ambas funciones si no fuera por el conjugado. Como nuestros datos no son funciones continuas, si no que los almacenamos como valores discretos, un conjunto de píxeles, debemos utilizar su versión discreta. Cambiamos la integral por un sumatorio:

$$(b * d)(\vec{x}) = \sum_{\vec{y}=-\infty}^{\vec{y}=\infty} b(\vec{y})^* d(\vec{x} + \vec{y}). \quad (11)$$

Lo importante de esto es que podemos utilizar el teorema de la convolución:

$$\mathcal{F}(b * d) = \mathcal{F}(b) \cdot \mathcal{F}(d). \quad (12)$$

Este teorema establece que, bajo determinadas circunstancias, la transformada de Fourier del producto de convolución de dos funciones es igual al producto de sus transformadas. Para calcular el producto de convolución podemos, usando este teorema, tomar la transformada inversa del producto de las transformadas.

³Planck Legacy Archive: <http://pla.esac.esa.int/pla/>

Tanto x como y son vectores, ya que representan un índice i, j en nuestras imágenes 2D. Tenemos entonces, que nuestro filtro adaptado genera un mapa al cual asigna a cada punto \vec{x} de la imagen, un valor de similitud con el perfil b en función del valor de los píxeles vecinos. Esto hace que tengamos que elegir un tamaño de las imágenes sobre las que vamos a filtrar. Este tamaño debe ser lo suficientemente grande para tener una buena estimación, pero que no aumente demasiado el tiempo de ejecución. Es por esto que no utilizamos el filtro sobre el mapa completo. Para calcular la correlación cruzada en una zona del cielo con el beam no necesitamos los píxeles que representan zonas del cielo enormemente distantes. Estos datos no influyen para el filtrado en esta zona, solo incrementan el número de datos y el tiempo de cálculo.

Sabemos que la fuente, al ser mucho más pequeña que el beam, toma la forma de este. Si aplicamos el filtro con la forma exacta del beam la amplitud del pico es invariante al filtrado. Como el ruido no toma la forma de beam ve su intensidad drásticamente reducida, ver imagen 6. Podemos decir ahora que si ponemos un umbral lo suficientemente por encima de la dispersión del ruido filtrado, todo lo que lo supere será una fuente compacta. Dependiendo de cómo de parecido sea nuestro perfil de filtrado al beam mayor será el SNR de las fuentes y más podremos bajar nuestro umbral, detectando así, fuentes más débiles.

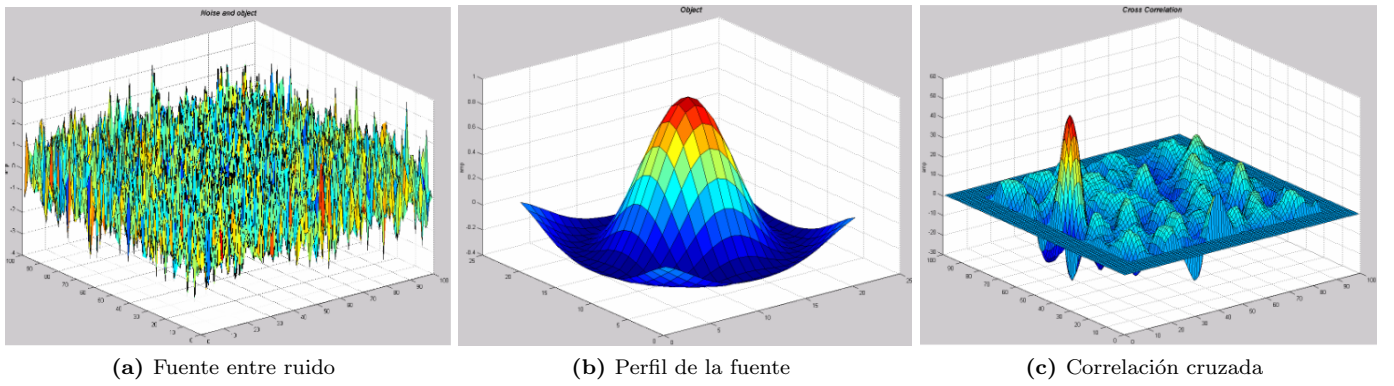


Figura 6: Ejemplo de correlación cruzada sobre un señal enterrada en ruido. Fuente: [16]

Antes mencionamos el teorema de la convolución y el uso de las transformadas en vez de la convolución. Esto se debe a que el algoritmo FFT (Fast Fourier Transform), transformada de Fourier rápida, nos permite realizar los cálculo de manera mucho más rápida. La FFT es un algoritmo que permite calcular la transformada de Fourier discreta (DFT) y su inversa. La DFT [17] se define como:

$$\mathcal{F}(b(k)) = \sum_{n=0}^{N-1} b(n) e^{-2ik\pi \frac{n}{N}}, k = 0, 1, \dots, N-1, \quad (13)$$

donde N es el número de datos (Píxeles en nuestras imágenes). La transformada inversa cambia el signo exponente y divide por N , esto hace que sea fácil de implementar a partir de la DFT. Tanto la convolución como la DFT requieren N^2 operaciones para calcularse, por eso no parcheábamos toda la esfera de una sola vez (también por que el ruido cambia en las diferentes zonas del cielo). La FFT requiere solo $N \log N$ operaciones. Cuando la cantidad de datos es grande la FFT supone una reducción enorme en el tiempo de computación. La FFT factoriza la matriz de la DFT en matrices con múltiples elementos nulos ya que es más rápido sumar que multiplicar. Para calcular la FFT en los bordes de una imagen el algoritmo genera datos simétrico falsos, los bordes de las imágenes filtrados no corresponden a la realidad, por lo que conviene no utilizar esos datos y realizar un nuevo filtrado sobre dicha zona. Aunque realicemos la misma operación sobre varios píxeles más de una vez sigue siendo más eficiente. Por otro lado al tratar con funciones discretas, se produce al realizar la transformada unas pequeñas fluctuaciones que dificultan la detección, haciendo que quizás una fuente que pasaría nuestro umbral no lo pase o viceversa, esto depende de la coordenada donde se aplique el parche y podría explicar diferencias en el número de detecciones con el mismo umbral por el mismo parche o similar.

La autocorrelación de un ruido $n(\vec{x})$ es una herramienta estadística que ayuda a encontrar patrones repetitivos de una señal. Se define como:

$$C(\vec{y}) = \langle n(\vec{x}) n(\vec{x} + \vec{y}) \rangle. \quad (14)$$

Es la correlación cruzada de una señal consigo misma. En el espacio de Fourier la autocorrelación del ruido es su espectro de potencias:

$$P(k) = \mathcal{F}(C)[m] = \sum_{n=1}^N C e^{-2ik\pi \frac{mn}{N}}, \quad m = 0, 1, 2, 3, \dots \quad (15)$$

En el espacio de Fourier el filtro adaptado (ψ) es proporcional al inverso del espectro de potencias del ruido. Tenemos entonces un problema, necesitamos conocer el espectro de potencias del ruido. No lo conocemos así que debemos obtenerlo de los datos. Si suponemos que las fuentes compactas son escasas en nuestros datos (si el trozo que filtramos es pequeño puede no ser el caso. La densidad de flujo de unas pocas fuentes puede dominar sobre esa región) podemos realizar una buena aproximación del espectro de potencias del ruido (sigue siendo una aproximación). Por otra parte la inversa de la transformada de Fourier del filtro es proporcional al perfil del filtro:

$$\mathcal{F}(\psi) \propto \frac{\tau}{P} \quad (16)$$

Esta relación de proporcionalidad nos indica que nuestro filtro ψ resalta datos que se parezcan a nuestro perfil τ y suaviza al ruido cuyo espectro de potencias es P . Conocemos nuestro perfil, es el beam de la antena. Falta explicar como se estima el espectro de potencias a partir de nuestros datos.

Uno de los métodos usados para calcular P , y el que utilizamos en este trabajo, es calcular la transformada de Fourier de nuestros datos, multiplicarla por su conjugado y promediar:

$$P(k) = \langle \tilde{d}(k) \tilde{d}(k') \rangle \delta(k - k'). \quad (17)$$

Promediamos de forma radial en anillos, un poco separados, alrededor del centro del parche. El problema es que tenemos una imagen pixelizada y en cada anillo tenemos unos pocos píxeles nada más (Si pilláramos parches mayores, pese a tener más datos, el ruido no sería homogéneo y no tendríamos una buena estimación de P). Uno de los parámetros del filtro es NBIN. NBIN gobierna sobre la separación entre los anillos. Si los radios están muy poco separados entre sí tenemos muchos paquetes (NBIN es grande), pero muy pocos píxeles y el promedio es muy malo. Si los radios están muy separados, tenemos pocos paquetes (NBIN es pequeño) y el promedio es mejor, pero perdemos detalle de como va la curva P en función del radio. Debemos crear un balance entre ambas partes, P bien estimada pero poco definida o conocer bien la curva pero teniendo una mala estimación de P . El parámetro NBIN, que influye en el tamaño de los paquetes, no estaba optimizado, por lo que una primera parte de este trabajo fue optimizarlo. Filtramos diez mil imágenes cuadradas de 512 píxeles de lado, con ruido blanco gaussiano (valores entre -1 y 1), completamente aleatorio y sin correlación espacial. En el centro de estas imágenes colocamos fuentes gaussianas, con una anchura a media altura de 8 píxeles. Un ejemplo de estas imágenes (salvo por el tamaño) es la figura 7. Cada imagen tenía un ruido aleatorio distinto.

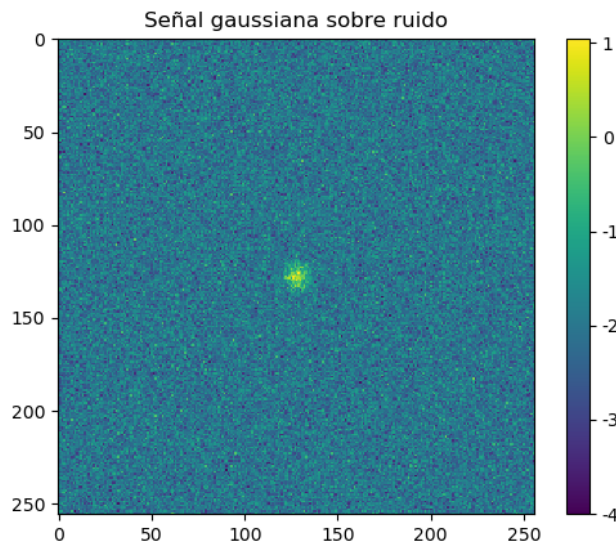


Figura 7: Señal gaussiana en el centro de una imagen de 256×256 píxeles con ruido blanco.

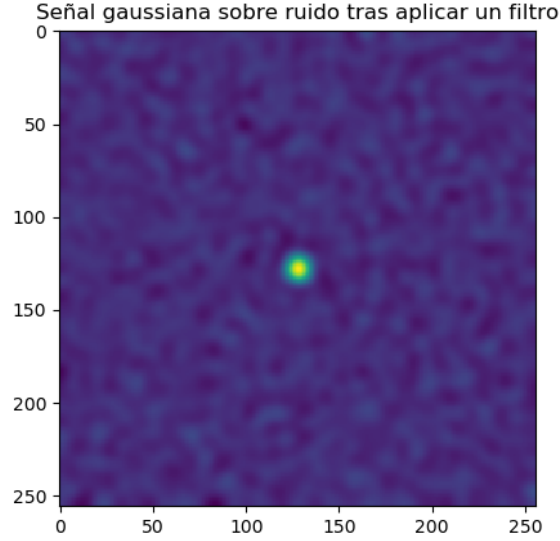


Figura 8: Misma imagen después de ser aplicado nuestro filtro adaptado.

Realizamos el filtrado para valores de NBIN desde 1 hasta $N/2$ en todas las imágenes. Llegamos a la conclusión, como se puede ver en la imagen 9, que el carácter que maximiza la ganancia es $NBIN = N/2$, donde N es el tamaño del lado de la imagen. Esto demostró que en este contexto el tamaño óptimo de los paquetes es de un píxel.

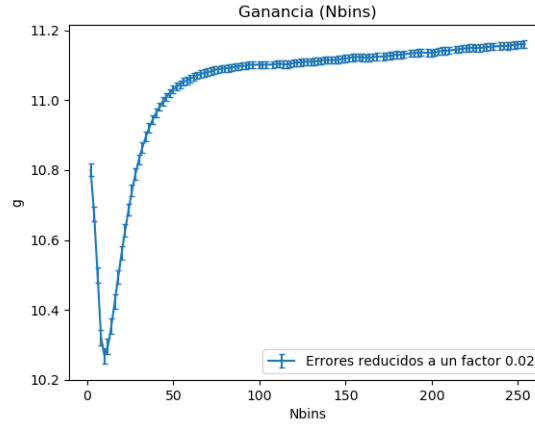


Figura 9: Ganancia de nuestro filtro sobre ruido blanco en función del parámetro NBIN.

Definimos la ganancia como:

$$g = \frac{SNR_f}{SNR} = \frac{A_f/\sigma_f}{A/\sigma}, \quad (18)$$

donde A es la amplitud de la fuente, σ es la dispersión del ruido en la imagen sin la fuente, y el sufijo $_f$ denota que la imagen ya ha sido filtrada. Como este filtro deja invariable la amplitud de la fuente, la expresión de la ganancia se reduce a:

$$g = \frac{\sigma}{\sigma_f} \quad (19)$$

Siempre y cuando $g > 1$ sabremos que el filtro está mejorando la imagen.

El FCM y el resto de emisiones de fondo en nuestras imágenes no son ruido blanco, si no ruido de color, ya que tienen una dependencia espacial. Queda probar el análisis de ganancia para el factor NBIN en las imágenes simuladas del FCM por Planck para las frecuencias del LFI, ver [18]. Estas imágenes fueron creadas a partir de simulaciones con los modelos teóricos de las fuentes en estas imágenes. Definimos tres regiones:

- REGION I: Una región de contaminación muy alta en latitudes de hasta 20 grados en torno al ecuador (por arriba y por abajo)
- REGION II: Una región de contaminación mediana entre los 20 y 60 grados por encima (o debajo) del ecuador
- REGION III: Una región de baja contaminación por encima de latitud 60° (y por debajo de latitud -60°)

Definimos estas regiones por que como se puede apreciar en la figura 5 el ruido de la galaxia tiene propiedades muy distintas en regiones cercanas al plano galáctico que lejos de este. El estudio del NBIN (que determina la calidad de la estimación del espectro de potencias del ruido) conviene hacerlo por separado en estas regiones.

Para estudiar el NBIN óptimo en estas regiones tomamos 192 parches de 256×256 píxeles, cubriendo unos $14,658^\circ \times 14,658^\circ$, de esta forma el tamaño de nuestros píxeles cubre un área similar a la de los píxeles de Healpy (Tamaño del píxel de Healpy: $4\pi/(12*1024*1024) = 9.986854e-07$ sr. Tamaño de nuestro píxel: $14.658^\circ*14.658^\circ/(256*256) = 9.986751e-07$ sr). Las posiciones de los parches son aleatorias. Con este número de parches, dado su área, cubrimos en mayor o menor medida toda la esfera. Realizamos el mismo procedimiento que antes: colocamos una fuente gaussiana en el centro, de anchura a media altura 8 píxeles. Con distintas intensidades, 2σ , 5σ y 10σ (σ es la desviación estándar del ruido en la imagen), débil, intermedia y brillante respectivamente. Filtramos cada imagen, con los tres tipos de fuentes en función de NBIN. Analizamos la ganancia para cada región

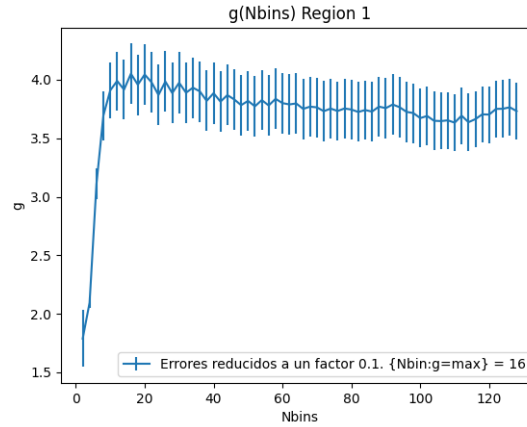


Figura 10: Ganancia de nuestro filtro sobre la emisión de la región 1 en función del parámetro NBIN.

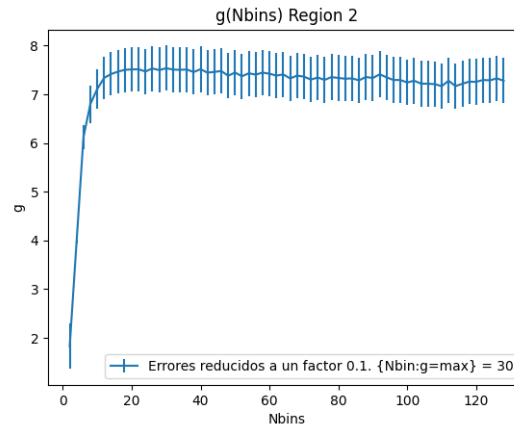


Figura 11: Ganancia de nuestro filtro sobre la emisión de la región 2 en función del parámetro NBIN.

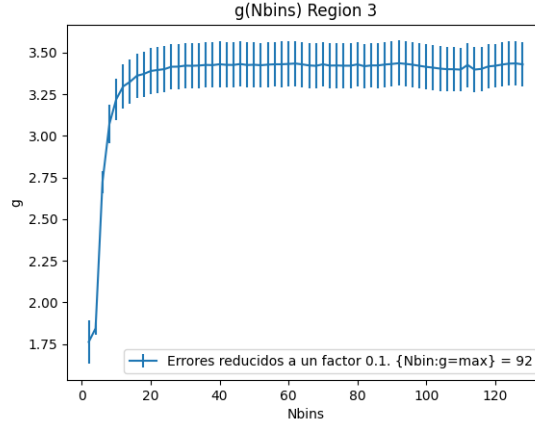


Figura 12: Ganancia de nuestro filtro sobre la emisión de la región 3 en función del parámetro NBIN.

Es interesante observar que, en las regiones 2 y 3, donde hay una menor contaminación al estar más alejados del plano galáctico, parece que la tendencia es a tener una ganancia máxima en $NBIN = N/2$. Esto mismo ocurría con el ruido blanco. Se observa mejor en la región 3 al ser la más alejada, aún así, la ganancia parece ser bastante constante para NBIN entre $N/4$ y $N/2$ en ambas regiones. En la región 1 se muestra un pico claro entorno a $NBIN = N/4$. Se trata claramente de un ruido de color proveniente de la Vía Láctea. Dado que para las regiones 2 y 3 la ganancia parece no influir demasiado para $NBIN = N/4$ o $N/2$ y que la ganancia es claramente superior para $NBIN = N/4$ en la región 1, se ha elegido este valor del parámetro para todo el trabajo. Esta decisión se basa en su simplicidad a la hora de programar el método y, ya que solo pretendemos hacer estimaciones del flujo en las fuentes detectadas y no requerimos de medidas de alta precisión en nuestro catálogo.

Es importante mencionar que si no pudiéramos conocer el espectro de potencias de nuestro ruido a partir de los datos, bien porque la imagen está parcialmente incompleta o cubierta por una gran cantidad de “malos” píxeles, pero aun así queremos utilizar una especie de detección utilizando un corte o umbral, podemos utilizar un filtro de de Maar o Mexican Hat Wavelet (MHW). Los filtros MHW convolucionan los datos con el núcleo de la ondícula y luego aplican el umbral de detección. La detección se optimiza variando un factor de escala a de manera que la desviación estándar de la imagen convolucionada es mínima. Existen multitud de variaciones del MHW, se las conoce como la familia de los MHW. Una comparación entre la eficiencia del filtrado simulaciones de Planck entre dos miembros de esta familia y el filtro adaptado se realizó en [19]. El método de detección de Planck utilizó el MHW2 un filtro optimizado que como se explica en [19] genera un rendimiento similar al del filtro aplicado y es sencillo de implementar. Al usar un filtro diferente es inevitable tener sutiles diferencias que apreciaremos al comparar nuestros datos con los de Planck.

4 Implementación del método de detección: Estructura y funcionamiento.

Una vez explicado los datos sobre los que se han trabajado y las herramientas utilizadas, resta explicar como funciona la implementación del método de detección en Python.

La idea del trabajo es generar una especie de caja negra, a la que aportando una serie de datos, entre los cuales están: el “path” (dirección dentro del disco duro en el cual esta alojado el archivo .fits con los datos), la frecuencia a la cual se ha tomado el mapa, la FWHM de la antena, y el umbral de detección, es decir, el SNR tras aplicar el filtro para considerar un pico de intensidad como una fuente compacta. Con estos datos se devuelve el catálogo de fuentes compactas. Con esta idea en mente se ha programado el código de la siguiente manera:

El programa está escrito en archivo .py, es decir un fichero de Python en la versión Python3, ya que Python2 está quedando en desuso y no se actualizan librerías o algunas rutinas nuevas ni se implementan. En la carpeta src, lugar donde está el código de nuestro método de detección, también se incluyen otros programas que reúnen varias rutinas

básicas. El uso de estas rutinas es muy habitual. Las escribimos en otro fichero que importamos. Podemos recurrir desde nuestro código a este fichero sintetizando varias líneas de código en una sin perder eficiencia. Dichas rutinas fueron gentilmente proporcionadas por el director del trabajo lo que se le agradece. El código incluye, además de estas rutinas antes mencionadas, varias rutinas de librerías de la comunidad como las anteriormente mencionadas Healpy y Astropy, de gran uso e importancia en la comunidad científica en el campo en que se desarrolla este trabajo. El funcionamiento del código será desarrollado a continuación paso a paso y mencionando cada herramienta utilizada.

Dentro de nuestro archivo .py que incorpora el método, se establece la función que filtra el FCM y genera los catálogos.

```
def filter_cmb(path="/home/chema/Escritorio/datos/LFI_SkyMap_030-BPassCorrected_1024_R3.00_full.fits", f=30, fwhm=32.65, cut=4, print_sources=False):
```

Esta función se puede llamar desde otro programa escrito en Python o desde la consola de Python ejecutandose desde la carpeta del src.

A continuación se cargan en la memoria del ordenador los datos aportados, asignándole las unidades correspondientes y guardándolos como variables para utilizar más adelante. Se crea una variable para guardar el tiempo en el que se inició en código, para así después saber el tiempo de ejecución total del programa. Una vez leído el mapa del FCM se cambia sus unidades a Jy, ya que suelen venir como K, representando la temperatura de un cuerpo negro que emitiría dicha radiación.

```
star_time = time.time()

# Loading of the FITS file with the image of the CMB and change the units to Jy. Specify the path
# to the file ,
# the frequency and the fwhm of the instrument.
f, fwhm = f * u.GHz, fwhm * u.arcmin # Input data

# The first element of the array is the one with the T map.
object_map = Fitsmap.from_file(path, freq=f, fwhm=fwhm)[0]

print("—— Data loaded ——")

# Changing data from K to Jy
object_map.to_Jy()
```

Ahora se carga en una variable el tamaño de los parches que proyectaremos sobre la imagen del FCM. Calculamos la posición de estos en el sistema de coordenadas galáctico de forma que cubran de manera equiespaciada toda la esfera y todos los píxeles al menos dos veces, de hecho el 98.44 % los observamos más de dos veces, evitando así que pudiéramos perder alguna fuente debido a que esta no se viera bien en un parche pero sí en otro, y más importante, debido a los efectos de borde creados al filtrar. De esta forma los píxeles que en un parche están en un borde, en otro no lo están y nos aseguramos de no perderlos.

```
# —— Parameters that characterize the patches that will be used to cover the CMB map on
# which we will use the filter ——
patch_size = 256

# —— Computing the center of the patches to cover all the sphere ——
nside = 8
npix = hp.nside2npix(nside)

centers = [hp.pix2vec(nside, i) for i in range(npix)]
# List with the position vectors of the npix of side = 8 that will be the center of the patches.

lon, lat = hp.vec2ang(np.array(centers), lonlat=True) # Center vector to lon and lat in deg.

npatches = len(centers)
```

Gracias a Healpy podemos representar el centro de los parches entorno a la proyección de la esfera en un plano, asegurándonos de que todo funciona correctamente, ver imagen 13.

Procedemos ahora a crear todos los parches entorno a las coordenadas previamente calculadas. Repitiendo para cada parche la misma operación: Filtrar, buscar máximos según el umbral que hemos pasado como parámetro a nuestra

función. Para cada máxima guardamos el valor del píxel central en J_y , como estimación del flujo de la fuente, y una estimación de su error como el valor del fondo entorno a la fuente. El valor del fondo lo tomamos en un anillo centrado en la fuente de radio interior igual a 3 veces la FWHM del aparato y radio exterior de 5 veces la FWHM. Esto es así porque si suponemos que la fuente es gaussiana, asunción que hemos tomado para filtrar, el valor de la señal a decaído lo suficiente como para que el ruido, la emisión de fondo, domine a la emisión de la fuente a distancias mayores de 3 veces la FWHM. Por encima de 5 veces la FWHM la emisión de fondo puede variar o incluso colarse emisión de otras fuentes. Estos valores que hemos guardado, así como sus coordenadas los guardamos en una tabla, es una primera versión de nuestro catálogo.

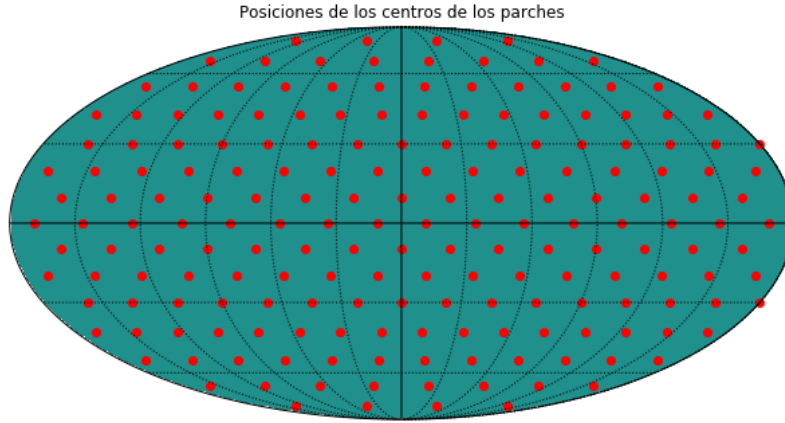


Figura 13: Ubicación de los centros de cada parche distribuidos de forma equiespaciada. Cada línea representa un intervalo de 30° .

```
# — Filtering every patch and getting all the candidates of compact sources, gathering all in a
# table —
peaks = []

for i in trange(npatches):
    coord = SkyCoord(l=lon[i], b=lat[i], frame='galactic', unit=u.deg)
    patch = object_map.patch(coord, npix=patch_size)
    filtered_patch = patch.matched(fwhm=object_map.fwhm)
    std = filtered_patch.std()
    threshold = cut * std
    indices = peak_local_max(filtered_patch.datos, min_distance=2,
                             threshold_abs=threshold, exclude_border=True, indices=True)
    for x, y in indices:
        coordinate = filtered_patch.pixel_coordinate(x, y)
        # Estimation of the flux error
        s = (object_map.fwhm/filtered_patch.pixsize).si.value
        w = np.arange(0, patch_size, 1, float)
        v = w[:, np.newaxis]
        r = np.sqrt((w - x) ** 2 + (v - y) ** 2)
        m = (r >= 3*s) & (r <= 5*s)

        peak = {'Longitude': coordinate.l.value,
                'Latitude': coordinate.b.value,
                'Flux': filtered_patch.datos[x, y],
                'Flux_error': filtered_patch.datos[m].std(),
                'SNR': filtered_patch.datos[x, y] / std}
        peaks.append(peak)

print("— Map filtered —")

table = Table(peaks)
```

La forma en que encontramos los máximos en los parches es gracias a Scikit, un paquete de Python diseñado para el análisis científico de imágenes. La idea para buscar un máximo local es sencilla: buscar para cada píxel el valor de sus vecinos. Si todos los vecinos son menores que el valor del propio píxel, estamos frente a un máximo. Esta tarea

es ardua y, pese a ser simple, molesta de programar. Además incluye problemas como que varios píxeles contiguos tengan valores idénticos y, su mayor defecto, para imágenes con una gran cantidad de píxeles puede requerir bastante tiempo. Afortunadamente Scikit cuenta con una rutina (`peak_local_max`) que permite realizar todo esto de una manera eficiente. Lo más importante es que nos permite excluir los bordes (sabemos que los datos aquí tras filtrar son falsos) y exigir una distancia mínima entre dos picos. (al menos dos píxeles) evitando fuentes extensas.

El problema ahora es el siguiente: Hemos parcheado la esfera de forma que cubrimos todo el cielo. Hay solapamientos entre los parches que permiten ver cada píxel una, dos o, como ocurre en la gran mayoría de las veces, más de dos veces. Esto, como se menciona antes, es muy útil. Los píxeles que tomamos en los bordes de un parche si tuvieran alguna fuente no la vamos a detectar. Por otra parte, aquí el problema, los píxeles que no caen en los bordes y contienen fuentes, harán que tengamos repeticiones en nuestro catálogo. Podríamos pensar que aunque tengamos varias veces la misma fuente en nuestro catálogo, como tienen la misma coordenada, sería fácil ubicar las copias y quedarnos solo con una. La realidad es más complicada que eso. Debido a que estamos proyectando un parche cuadrado sobre una superficie esférica, estamos distorsionando ligeramente las imágenes. Esto es similar a como un mapamundi, tan familiar para todos, distorsiona el tamaño de los países. Parece que Groelandia es del tamaño de África cuando en realidad es unas 13 veces menor. De igual manera los píxeles más alejados del centro de los parches se ven más distorsionados. Esto hace que al transformar la coordenada de un píxel i,j en un parche, a coordenadas galácticas l,b (rutina de Astropy) en un parche o en otro obtengamos unas sutiles diferencias, haciendo que, pese a ser la misma fuente, no caigan en la misma coordenada en nuestro catálogo. Por otra parte la rutina que transforma coordenadas al hacer los cálculos redondea a una determinada cifra decimal, con lo que se introduce otro ligero error. Como las repeticiones no caen en las mismas coordenadas, no nos queda más remedio que calcular las distancias angulares de una fuente con el resto. Utilizamos las coordenadas para saber las distancias sobre la superficie de la esfera. Sabemos que las perturbaciones, al ser los parches relativamente pequeños, no serán muy grandes, al menos no mayores que la FWHM. Calcular las distancias de una fuente con el resto para cada fuente del catálogo y, buscar aquellas a distancias menores que una dada, es una tarea difícil de programar. Usando las tradicionales búsquedas mediante bucles for, las cuales son altamente ineficientes en Python (lenguaje Interpretado y no compilado), llevaría horas para catálogos con unas pocas miles fuentes, del tamaño que manejamos. Afortunadamente Astropy incorpora dos rutinas: Una para calcular distancias sobre la superficie de la esferas (`separation`). La otra realiza la búsqueda en dos catálogos de fuentes cercanas (`search_around_sky`). Esta última incorpora a la primera, solo necesita una lista con las coordenadas de las tablas a cotejar. El método `search_around_sky` utiliza un algoritmo de búsqueda de tipo árbol binario, mucho más eficiente y fuera del nivel de un alumnado general de cuarto curso de grado en física (afortunadamente ya esta implementado). Este método esta pensado para cotejar dos catálogos distintos, pero utilizando como datos de entrada las coordenadas de un catálogo dos veces, permite hacer búsquedas internas en ese catálogo.

Una vez que sabemos como ubicar las repeticiones quedaría quedarnos con una de las repeticiones. Debido a la distorsión los valores del flujo y las coordenadas varían, necesitaríamos saber cual es la mejor estimación. Además puede que una fuente B este en el radio de búsqueda de una fuente A y otra C, mientras que A y C no coinciden en nuestros radios de búsqueda. En un primer momento no sabemos si B pertenece a A o a C. Las versiones anteriores solventaron este problema aplicando unos criterios para descartar una fuente u otra, sin embargo, una idea original llevada a cabo en este trabajo para descartar las repeticiones, ha sido la siguiente:

`Search_around_sky` nos devuelve entre otras cosas dos listas de índices: los índices de las fuentes, cuyas coordenadas fueron pasadas como parámetro 1, en su tabla y, las fuentes, cuyas coordenadas fueron pasadas como parámetro 2, que están dentro del rango de búsqueda de la fuente en la tabla 1. Se podría visualizar de la siguiente forma:

```
[1, 1, 2, 3, 4, 5, 5, ... , 200, 200]
[1, 5, 2, 3, 4, 1, 5, ... , 184, 200]
```

Es decir, si no encuentra ninguna coincidencia, solo a sí misma tendremos el índice una vez en cada lista. Si hay coincidencias lo tendremos repetido en la primera lista y los valores de coincidencia en la segunda.

Sabiendo como son las listas que nos devuelve con las coincidencias, generamos una tabla de correspondencias con estos índices y sus distancias.

```
coordinates = SkyCoord(l=np.array(table['Longitude'].tolist()),
                        b=np.array(table['Latitude'].tolist()),
                        frame='galactic',
                        unit=u.deg) # Coordinates of each peak in the galactic system (in angles)
```

```

# — Searching for repetitions in our catalogue —

resolution_max = object_map.fwhm.to(u.arcmin)

ind1, ind2, dist, d3d = search_around_sky(coordinates, coordinates, 3/4*resolution_max)

correspondence_table = Table()
correspondence_table['Index 1'] = ind1
correspondence_table['Index 2'] = ind2
correspondence_table['Separation'] = dist

final_table = Table()

```

La distancia que hemos usado es $3/4$ de la FWHM. Dos fuentes muy cercanas, las cuales podemos discernir, pueden estar a una distancia menor que la FWHM debido a la distorsión. Podemos hacer esto sabiendo que la distorsión en la coordenada es bastante menor que la FWHM en las frecuencias trabajadas.

El tamaño de la tabla de correspondencias es mayor o igual que nuestro catálogo (igual en el caso de que no hubiera repeticiones). Hacemos un bucle iterativo condicionado. Mientras la tabla de correspondencias sea mayor que nuestro catálogo, significará que hay repeticiones. Recorriendo cada índice en el catálogo: Tomamos una tabla con las repeticiones para cada índice. Si el tamaño de esta tabla es de sólo una fila, significa que no hay repeticiones, por lo que copiamos sus datos en una nueva tabla. Si, por otra parte, es mayor que uno, hay repeticiones. Cuando hay repeticiones tomamos un parche centrado entorno a esa fuente. Filtramos y añadimos los datos de los máximos en un área pequeña entorno a nuestra fuente, justo del tamaño de nuestra búsqueda de repeticiones, $3/4$ de la FWHM. Añadimos a una lista de valores que ignoramos, los índices de la lista dos que coinciden con el índice en el que estamos. Tras todo esto generamos de nuevo una búsqueda de coincidencias en nuestro catálogo provisional. Sería de esperar que tras una iteración en este bucle, se eliminaran todas las repeticiones. Se puede encontrar que al retomar una fuente antes distorsionada (coordenadas ligeramente cambiadas), que no entraba en el área de búsqueda de otra fuente, ahora con las coordenadas buenas si que entre. Con una par de iteraciones se descartan rápidamente todas las repeticiones. Cada iteración requiere menor tiempo y recursos que la iteración previa al tener que descartar menos repeticiones y tratar con listas más pequeñas.

```

# — Resolution of coincidences —
while len(correspondence_table) > len(final_table):
    skipped = []
    final_table = Table()
    for i in range(len(table)):
        if i == 0:
            final_table = Table(rows=table[i], names=('Longitude', 'Latitude', 'Flux', '
Flux_error', 'SNR'))
        if i in skipped:
            continue
        x = correspondence_table[correspondence_table['Index 1'] == i]
        if len(x) == 1:
            if i == 0:
                continue
            else:
                final_table.add_row([table['Longitude'][i], table['Latitude'][i],
                                     table['Flux'][i], table['Flux_error'][i],
                                     table['SNR'][i]])
        else:
            skipped = np.append(skipped, np.array(x['Index 2'].tolist()))
            coord = SkyCoord(l=table['Longitude'].tolist()[i], b=table['Latitude'].tolist()[i],
                             frame='galactic', unit=u.deg)
            patch = object_map.patch(coord, npix=int(patch_size/2), deltatheta_deg=14.658/2)
            filtered_patch = patch.matched(fwhm=object_map.fwhm)
            std = filtered_patch.std()
            threshold = cut * std
            indices = peak_local_max(filtered_patch.datos, min_distance=2,
                                     threshold_abs=threshold, exclude_border=True, indices=True)

    for x, y in indices:
        s = (object_map.fwhm / filtered_patch.pixsize).si.value
        if (x >= patch_size/4 - 3/4*s) & (x <= patch_size/4 + 3/4*s) & \
            (y >= patch_size/4 - 3/4*s) & (y <= patch_size/4 + 3/4*s):

```

```

coordinate = patch.pixel_coordinate(x, y)

# Estimation of the flux error
w = np.arange(0, int(patch_size/2), 1, float)
v = w[:, np.newaxis]
r = np.sqrt((w - x) ** 2 + (v - y) ** 2)
m = (r >= 3 * s) & (r <= 5 * s)

peak = {'Longitude': coordinate.l.value,
        'Latitude': coordinate.b.value,
        'Flux': filtered_patch.datos[x, y],
        'Flux_error': filtered_patch.datos[m].std(),
        'SNR': filtered_patch.datos[x, y] / std}
final_table.add_row([peak['Longitude'], peak['Latitude'], peak['Flux'],
                    peak['Flux_error'], peak['SNR']])

if i == 0:
    final_table.remove_row(0)

coordinates = SkyCoord(l=np.array(final_table['Longitude'].tolist()),
                       b=np.array(final_table['Latitude'].tolist()),
                       frame='galactic',
                       unit=u.deg)
ind1, ind2, dist, d3d = search_around_sky(coordinates, coordinates, 3/4*resolution_max)

correspondence_table = Table()
correspondence_table['Index 1'] = ind1
correspondence_table['Index 2'] = ind2
correspondence_table['Separation'] = dist
table = final_table

print("—— Duplicated sources removed ——")

```

Lo interesante de este nuevo método es que permite, de una sola vez, resolver todas las coincidencias de una fuente. Ahorra lidiar con si la mencionada fuente B pertenece a la fuente A o C. Además de ser bastante eficiente permite, por otra parte, saber que el valor de las fuentes al resolver una repetición es el mejor posible (mínima distorsión, fuente en el centro) con nuestra estimación.

Tenemos ahora una lista final de fuentes compactas detectadas mediante nuestro filtro adaptado, sin repeticiones. Incluye sus coordenadas en el sistema galáctico, una estimación de su flujo en Jy junto a su error y, su SNR. Queda guardar dicho catálogo en una tabla en un fichero .fits. Añadir antes dos columnas con sus coordenadas en el sistema ecuatorial (ascensión recta y declinación), por comodidad para el futuro usuario de este catálogo. Además si al llamar a la función se indicó, `print_sources=True`, genera una imagen de la proyección de la esfera con puntos señalando la posición de las fuentes. Esta opción está deshabilitada por defecto. Finalmente muestra en pantalla el tiempo de ejecución.

```

# Saving a picture with the localization of all the sources.
if print_sources:
    hp.projscatter(table['Longitude'].data.data,
                  table['Latitude'].data.data,
                  lonlat=True, coord='G', color='r', marker='.')

    hp.graticule()
    hp.projscatter(table['Longitude'].data.data,
                  table['Latitude'].data.data,
                  lonlat=True, coord='G', color='r', marker='.')

    plt.savefig('Detected_Sources_' + str(f) + '.png')

print(str(len(table)) + ' compact sources detected.')
print('Execution time: ' + str(time.strftime("%M:%S", time.gmtime(time.time() - star_time))) +
      ' (min:s)')

```

El código va pintando en la consola de Python, o en la del compilador de texto que uses, frases indicando los puntos que se están realizando, barras de porcentaje mientras filtra y elimina repeticiones, dando una idea del tiempo que le falta a dicho paso. Al terminar indica el tiempo final que ha tardado en aplicar el filtro y obtener el catálogo, así como el número de fuentes detectadas.

Un ejemplo de los mapas con la ubicación de las fuentes detectadas y las tablas puede verse en las figuras 14 y 15.

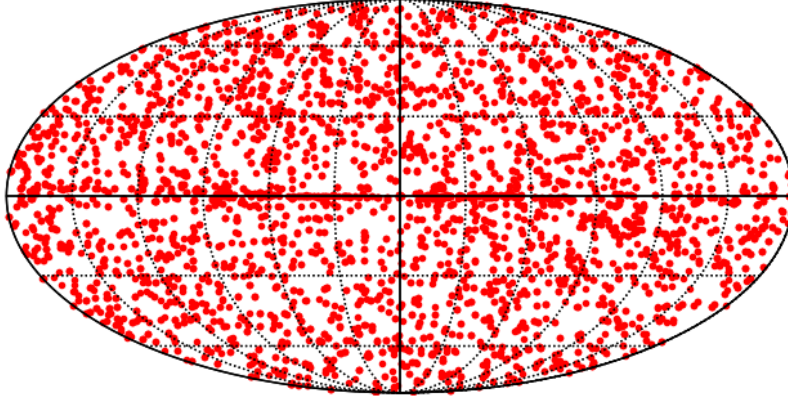


Figura 14: Mapa indicando la posición de cada fuente detectada, utilizando nuestro método, en las imágenes del FCM proporcionadas por Planck en el año 2015, a la frecuencia del LFI de 44 GHz. Cada línea representa un intervalo de 30° .

Longitude deg	Latitude deg	Flux Jy	Flux_error Jy
86.0204081632653	83.33765961775218	2.7908605670828086	0.10404754850949924
56.69117647058823	80.66988136332392	2.098585578878311	0.10823482224566032
3.713592233009709	80.52094614691819	0.9736236690877766	0.09117780513190893
23.000735967757272	80.9785531883568	0.8997982058720706	0.10407049381364253
64.82315942337326	80.36022451599425	0.6630991736567226	0.14052033263377517
37.20297029702971	80.81881059506735	0.6242011810215335	0.14165975949774545
87.08955223880596	80.80738629522925	0.6045608821481214	0.2503182657189907
101.65816220257582	80.6582861155751	0.6913772785250561	0.104717254144621
232.81135300406754	84.89274419503478	0.6705781384338447	0.09546547366596916
165.63829787234044	79.24817170667029	0.5560117638568853	0.07614856271387736
...
276.73635489360015	-80.10766582870308	0.5669076492396071	0.08216284419067876
249.40647059997818	-83.43044636378389	0.3490358074615892	0.11145997851662681
244.84048189555648	-71.94804281385687	0.3355081836519848	0.08804117035800081
335.1650943396226	-80.30313516468107	0.8492132808314661	0.08460230615936655
98.59560517508675	-87.94325171318651	0.996435633477404	0.10967058973225714
68.0625	-86.28729351960897	0.7744886446313436	0.1261687079434113
4.433083945556305	-82.25842761682782	0.3362848980177356	0.12152216132033111
76.68994519255573	-87.1711674090818	0.3342468567435698	0.09545304992281575
216.84910322589508	-86.61830857909513	0.7558895820740861	0.08656107983160842
171.4875341574107	-85.7956147504512	0.5853316781790806	0.09134053292936598

Figura 15: Catálogo obtenido a partir del filtrado de las imágenes del FCM proporcionadas por Planck en el año 2015, a la frecuencia del LFI de 30 GHz. Se muestran en este caso las longitudes y latitudes galácticas, una estimación del flujo y su error.

5 Presentación y análisis de resultados

5.1 Presentación de resultado y comparación con los PCCS2

Una vez explicado el método de detección, así como su implementación en Python, queda hablar de los resultados obtenidos. Debemos compararlos con los datos obtenidos en las versiones anteriores del método, analizarlos, y sacar conclusiones de la efectividad y rendimiento de esta nueva versión.

Se ha aplicado el código creado sobre los mapas de Planck del LFI (30, 44 y 70 GHz) publicados en 2015. El umbral de detección impuesto fue de un SNR mínimo de 3,5. Esto ha sido posible debido a que el filtro utilizado ha demostrado funcionar razonablemente bien. Cuanto mejor es el filtro más reduce el valor del ruido, lo que permite imponer un umbral de detección mayor (sin riesgo de detectar fluctuaciones del ruido filtrado como fuentes). El valor estándar utilizado como umbral de detección es el conocido 5σ , es decir un SNR de 5 (un valor mínimo de intensidad igual a cinco veces la dispersión estándar del fondo).

Antes de analizar los datos es importante mencionar que pese a que las imágenes del FCM tomadas por Planck indiquen que han sido obtenidas a las mono-frecuencias de 30, 44 y 70 GHz, esto no es del todo correcto. El experimento LFI no dispone de una antena que detecte a una mono-frecuencia, por el contrario, se trata de un array (un conjunto de radio antenas) con una curva de sensibilidad a diferentes frecuencias. Su curva de respuesta no es una delta centrada en las frecuencias mencionadas, sino que se trata de una curva más o menos fina entorno a dichas frecuencias. Detecta las frecuencias de 30, 44 y 70 GHz más frecuencias ligeramente superiores e inferiores. El tamaño del beam (la respuesta angular de la antena) que considerábamos gaussiano con una FWHM dada, se deforma ligeramente y su FWHM cambia. Para saber más sobre la calibración del LFI y como es realmente el beam, ver [20]. Esto afecta ligeramente al filtro, variando la detección de fuentes compactas. Afecta también a como pasamos los datos de K a Jy, por lo que induce errores en la estimación de los flujos. Una corrección completa que incluya todos estos detalles se escapa al nivel de este trabajo de fin de grado, afortunadamente hay unas frecuencias y FWHM efectivas ligeramente distintas a las frecuencias del LFI. Estos parámetros efectivos vienen dentro de un conjunto de datos conocidos como RIMO, Reduced Instrument MOdel:

- Frecuencias [30 ; 44 ; 70] GHz \rightarrow [28,4 ; 44,1 ; 70,4] GHz
- FWHM [32,65 ; 27,00 ; 13,01] arcmin \rightarrow [32,293 ; 27,00 ; 13,213] arcmin

Introducir estas frecuencias y FWHM efectivas consigue corregir, ligeramente, los efectos anteriormente descritos, ver figura 16.

Para ver estas diferencias escogemos del catálogo de fuentes compactas de Planck. Una cantidad de 20 fuentes a latitudes galácticas superiores a 45° (en valor absoluto) para evitar contaminación de la galaxia y, con intensidades comprendidas entre 2 y 10 Jy, de forma que sean lo suficientemente brillantes para observarlas y que no saturen la imagen. Una vez seleccionadas las fuentes parcheamos en esas coordenadas y filtramos. Cogemos el valor de intensidad como el valor en el píxel central tras filtrar. En el caso en que nuestras estimaciones fueran idénticas a las de Planck, tendríamos una recta de pendiente la unidad, con ordenada en el origen igual a cero. Lo normal es que no sea así, ya que hay fluctuaciones provocadas al parchear y filtrar, además de que no estamos usando el mismo filtro y el catálogo de Planck está más elaborado. Sin embargo esperamos obtener algo similar, una recta de distinta pendiente pero similar a la unidad. Dibujando la recta $y = x$ sobre el conjunto de puntos del valor de las intensidades para las misma fuentes con corrección y sin corrección podemos ver cual es mejor.

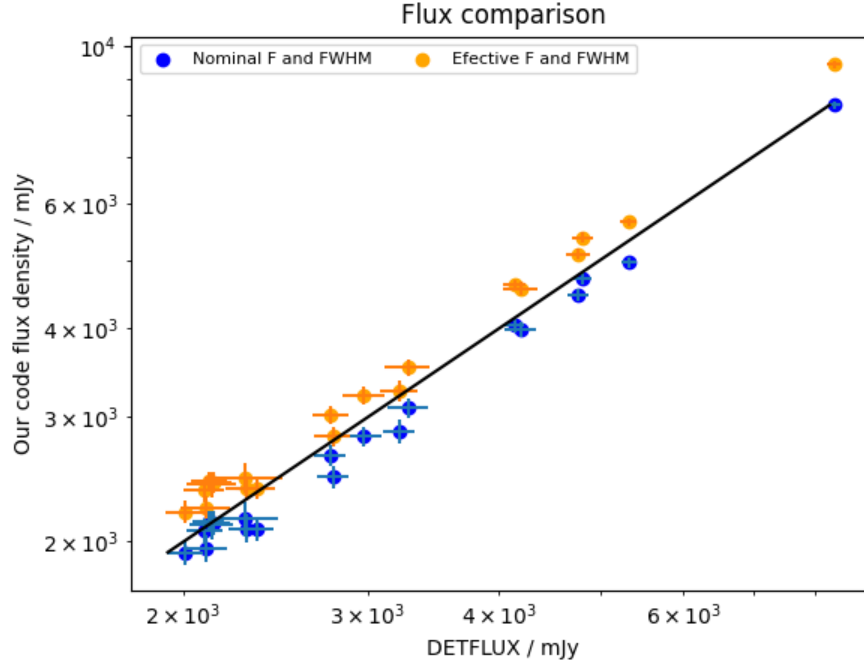


Figura 16: Comparación de los flujos obtenidos por Planck y por nuestro método para las mismas fuentes filtrando con los valores de frecuencias y FWHM nominales y efectivos. Estas pruebas se han realizado sobre el catálogo de fuentes compactas publicado por la misión Planck en 2015, de su experimento LFI, a la frecuencia de 30 GHz.

Podemos observar dos cosas de esta comparación. En primer lugar vemos que detectamos las todas las fuentes en ambos casos. De aquí deducimos que usar los parámetros nominales o efectivos no afecta gravemente a la posición de los máximos. Por otro lado vemos que sino aplicamos el cambio tenemos un error sistemático al estimar el flujo. En el caso del mapa tomado a 30 GHz (valor nomina), si no usamos los valores efectivos, estamos sobrestimando de manera sistemática el valor del flujo. Dependiendo de la imagen que usemos, la diferencia entre usar los valores nominales o efectivos varía. Queda establecido que hay un error al utilizar los valores nominales que se puede paliar, en cierto modo, utilizando los valores efectivos. Aunque se podría realizar un trabajo de filtrado más extenso, teniendo en cuenta la forma exacta del beam del LFI y la ventana de frecuencias del detector, dicho trabajo se escapa de la mira de este trabajo que busca generar a partir de estos mapas catálogos con aproximaciones del flujo de estas fuentes.

Una vez aplicado el filtro, utilizando los valores de frecuencia y FWHM del RIMO, obtenemos tres catálogos de fuentes compactas, uno por cada mapa de 30, 44 y 70 GHz. Cada catálogo incluye información de las fuentes detectadas como: sus coordenadas en los sistemas galáctico y ecuatorial (los dos más utilizados en este campo) el valor del flujo junto a su error en janskys, necesarias para otros estudios como, por ejemplo, la distancia de la fuente a nosotros. Finalmente incluye un valor del SNR que indica cómo de brillante es esa fuente en su entorno (mayor SNR más destaca entre la radiación de fondo).

- Catálogo 30 GHz:

Hemos detectado 1849 fuentes compactas frente a las 1560 fuentes de Planck. Una imagen de la ubicación de las fuentes puede observarse en la figura 17.

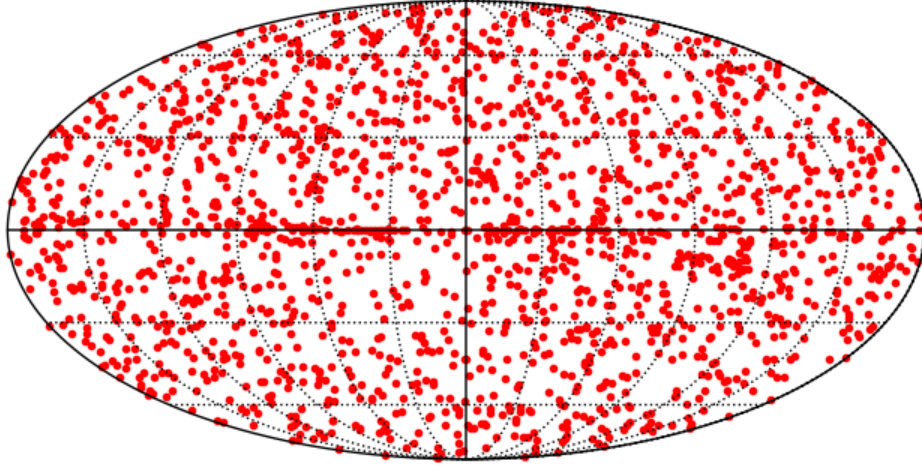


Figura 17: Ubicación de la fuentes compactas detectadas sobre las imágenes publicadas en 2015 por Planck del LFI tomadas a 30 GHz. Cada línea representa un intervalo de 30° .

Podemos observar que las fuentes están distribuidas de manera homogénea por la esfera celeste, habiendo una pequeña sobre densidad de estas en el plano galáctico, y en determinados lugares como las nubes de Magallanes. Por todo esto obtenemos un resultado satisfactorio (esperábamos algo similar). Queda, no obstante, comparar nuestro catálogo con el de Planck, un catálogo corroborado del cual podemos fiarnos. Si nuestras fuentes coinciden podremos asegurarnos de que el catálogo está aparentemente bien y que nuestro método parece funcionar.

De igual forma que buscábamos repeticiones en nuestros catálogos, cotejamos ahora el nuestro con el de Planck. Imponiendo un radio de búsqueda de $3/4$ la FWHM encontramos que 1296 de las 1849 fuentes de nuestro catálogo coinciden con fuentes de Planck. El resultado es prometedor. Queda comparar los flujos en nuestro catálogo con los del catálogo de Planck para estas fuentes. Representamos la comparación de los flujos, figura 18.

Observamos que el error en la estimación del flujo aumenta considerablemente cuanto más débil es la fuente. Parece que nuestra estimación del flujo coincide, dentro de los errores, con la de Planck, al menos en la mayoría de las fuentes detectadas. Hay que comentar que la forma de tomar los flujos de Planck no es como la nuestra, hemos utilizado filtros ligeramente distintos y pequeñas fluctuaciones se han dado al proyectar. Todo esto nos asegura que hay un margen de error entre las estimaciones de nuestros flujos. Observando la figura 18, podemos deducir con un alto grado de seguridad que hemos detectado 1296 fuentes que Planck también detectó.

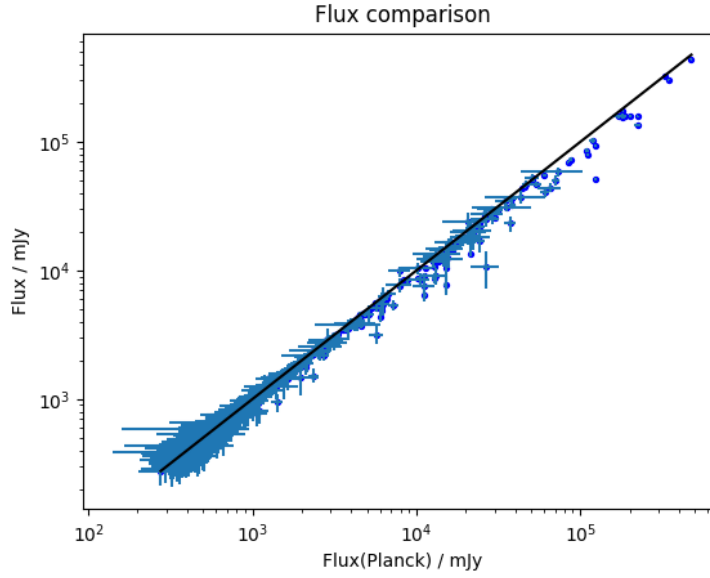


Figura 18: Comparación de los flujos para las fuentes de nuestro catálogo, obtenido al aplicar un filtro de 3.5σ de corte, con el catálogo de Planck. Mapa del FCM utilizado: LFI 30 GHz publicado en 2015.

- Catálogo 44 GHz:

Hemos detectado 1708 fuentes compactas frente a las 934 fuentes de Planck. Una imagen de la ubicación de las fuentes puede observarse en la figura 19.

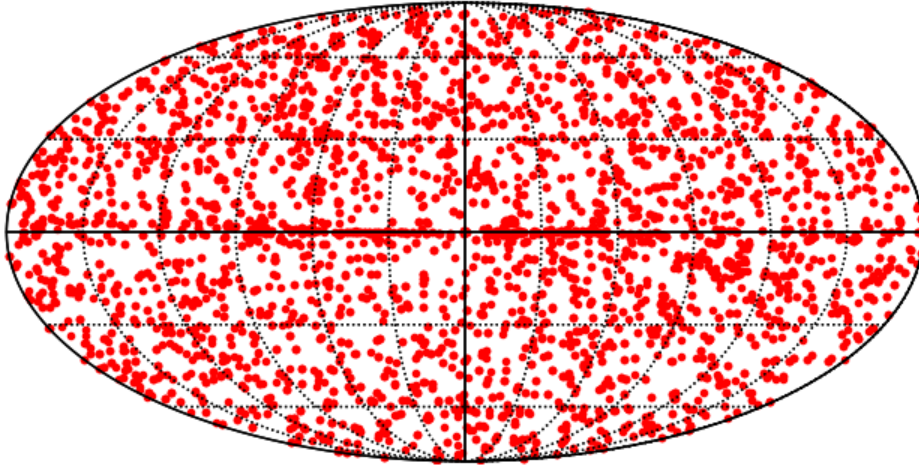


Figura 19: Ubicación de la fuentes compactas detectadas sobre las imágenes publicadas en 2015 por Planck del LFI tomadas a 44 GHz. Cada línea representa un intervalo de 30° .

De nuevo observamos que las fuentes detectadas están repartidas de forma homogénea por todo el cielo. Aparecen las ya mencionadas sobredensidades en el plano galáctico y en pequeñas zonas puntuales. De las 934 fuentes del catálogo de Planck tenemos 794 coincidencias. Veamos la comparación de los flujos.

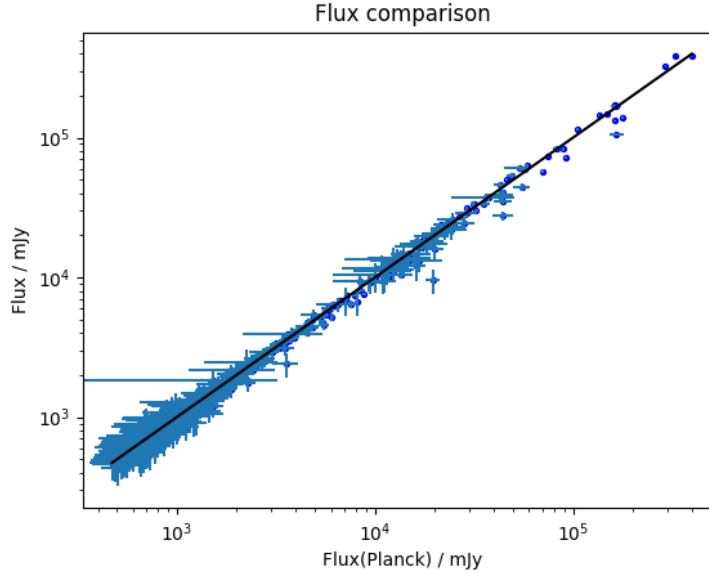


Figura 20: Comparación de los flujos para las fuentes de nuestro catálogo, obtenido al aplicar un filtro de 3.5σ de corte, con el catálogo de Planck. Mapa del FCM utilizado: LFI 44 GHz publicado en 2015.

Parece que las fuentes que encontramos en ambos catálogos son las mismas (dentro de sus márgenes de error). De igual forma se puede observar que las fuentes más débiles tienen mayor error que aquellas más brillantes. Cabe mencionar que el beam del LFI a 44 GHz es el que más cambia, por lo que esperamos que sea aquí donde más diferencias encontremos respecto a los catálogos de Planck.

La diferencia de fuentes encontradas respecto a Planck aumenta notablemente en comparación a los que veíamos a 30 GHz.

Queda comparar nuestro catálogo obtenido a 70 GHz con el de Planck.

- Catálogo 70 GHz:

Hemos detectado 4350 fuentes compactas frente a las 1296 fuentes de Planck. Una imagen de la ubicación de las fuentes puede observarse en la figura 21. Encontramos 1102 coincidencias entre ambos catálogos.

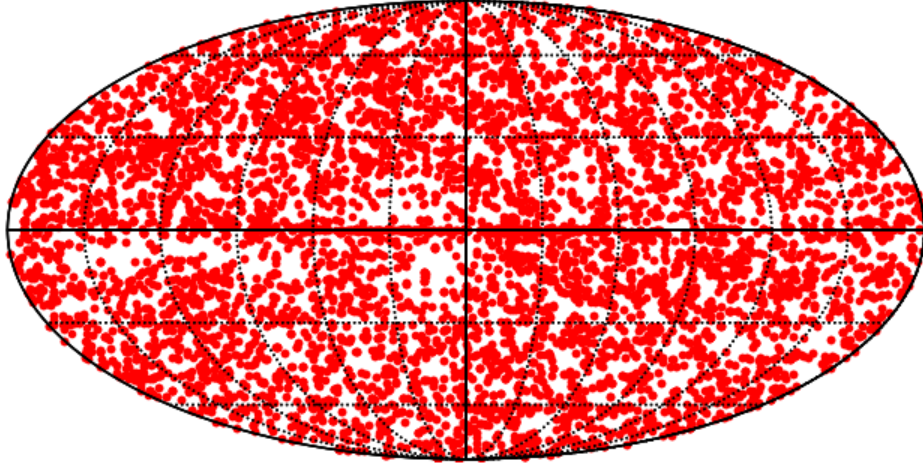


Figura 21: Ubicación de la fuentes compactas detectadas sobre las imágenes publicadas en 2015 por Planck del LFI tomadas a 70 GHz. Cada línea representa un intervalo de 30° .

Podemos nuevamente observar la dispersión homogénea de nuestras fuentes por la esfera celeste. Aparecen nuevamente las ya mencionadas sobredensidades de fuentes tanto en el plano galáctico, como en zonas puntuales (caso de las antes señaladas nubes de Magallanes).

Comparamos ahora el flujo de nuestras fuentes frente a las de Planck, para corroborar que son las mismas:

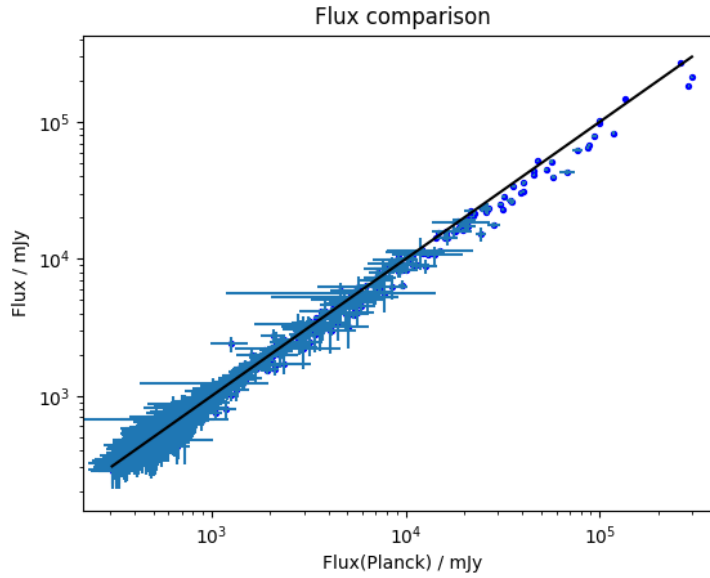


Figura 22: Comparación de los flujos para las fuentes de nuestro catálogo, obtenido al aplicar un filtro de 3.5σ de corte, con el catálogo de Planck. Mapa del FCM utilizado: LFI 70 GHz publicado en 2015.

Nuevamente, los resultados son satisfactorios. Podemos hablar, dentro de un grado de confianza significativo, de las mismas fuentes. Podemos afirmar que nuestro método funciona y que es capaz de detectar fuentes compactas a partir de imágenes del FCM.

Notamos por otro lado que, al aumentar la frecuencia de exposición a la que se tomaron los mapas de FCM que hemos filtrado, aumenta la cantidad de fuentes que detectamos. Esto es debido, mayormente, a que al aumentar la frecuencia el límite de resolución disminuye. De esta forma dos fuentes que antes estaban demasiado juntas y que no resolveríamos, tomando como una solo fuente, ahora las vemos claramente separadas. Esto no parece ser el caso en el catálogo de 44 GHz, se debe principalmente a que en esta frecuencia, el RIMO no es una buena aproximación, al menos peor que en los otros dos casos del LFI, 30 y 44 GHz.

Queda explicar las diferencias con los catálogos de Planck. Explicar que ocurre con las fuentes detectadas por Planck pero no por nosotros y viceversa. Para este análisis vamos a comparar nuestros catálogos, obtenidos al aplicar el filtro con un umbral de detección de $\text{SNR} = 4$ (el mismo que se usó en Planck. Puede demostrarse rápidamente mirando el mínimo SNR en dicho catálogo) con los PCCS2.

Los catálogos de Planck que hemos usado para comprar son los PCCS2, Planck Second Catalogue of Compact Sources, de 30, 44 y 70 GHz. Estos catálogos se publicaron en el año 2015 y se pueden obtener de la página web del Planck Legacy Archive⁴, donde se encuentran todos los datos obtenidos por la misión en sus diferentes experimentos.

Para saber más sobre los PCCS2 y como se obtuvieron ver [21]. Comparamos ahora los catálogos de Planck, PCCS2, con los nuestros usando el umbral de detección $\text{SNR} = 4$, igual que hizo Planck.

- Catálogo 30 GHz:

Encontramos en este caso 1378 fuentes frente a las 1560 de Planck. Esperábamos encontrar menos que en el caso en que usábamos un umbral de detección menor. Puede verse la localización de la fuentes en la imagen 23.

Observamos que las fuentes siguen distribuidas de la misma forma que con el umbral de detección anterior pese a ser menos. Están distribuidas de manera homogénea por la esfera, salvando el plano galáctico y unos pocos lugares.

De las 1378 fuentes detectadas tenemos 1152 coincidencias con el catálogo PCCS2 de Planck. Se asemeja bastante a el resultado en el caso anterior. Un análisis comparando flujos entre ambas catálogos será llevado a cabo más adelante.

- Catálogo 44 GHz:

Encontramos en este caso 991 fuentes frente a las 934 de Planck. Menos que en el caso en que usábamos un umbral de detección menor. Puede verse la localización de la fuentes en la imagen 24.

Las fuentes siguen distribuidas de la misma forma que con el umbral de detección anterior. Distribuidas de manera homogénea por la esfera salvo en el plano galáctico y unos pocos lugares.

De las 1708 fuentes detectadas tenemos 699 coincidencias con el catálogo PCCS2 de Planck. Nuevamente en gran concordancia con el caso anterior. Compararemos los flujos más tarde.

- Catálogo 70 GHz:

Encontramos en este caso 1845 fuentes frente a las 1296 de Planck. Puede verse la localización de la fuentes en la imagen 25.

Las fuentes siguen distribuidas de forma homogéneas salvo en el plano galáctico y unas pocas zonas.

De las 1708 fuentes detectadas tenemos 991 coincidencias con el catálogo PCCS2 de Planck. Compatible con el caso anterior. Compararemos los flujos a continuación.

⁴<http://pla.esac.esa.int/pla/#catalogues>.

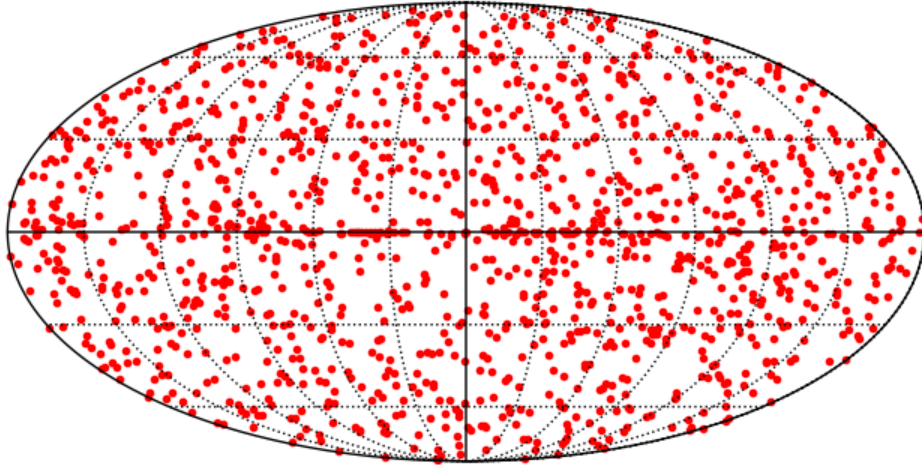


Figura 23: Ubicación de la fuentes compactas detectadas sobre las imágenes publicadas en 2015 por Planck del LFI tomadas a 30 GHz. Cada línea representa un intervalo de 30° .

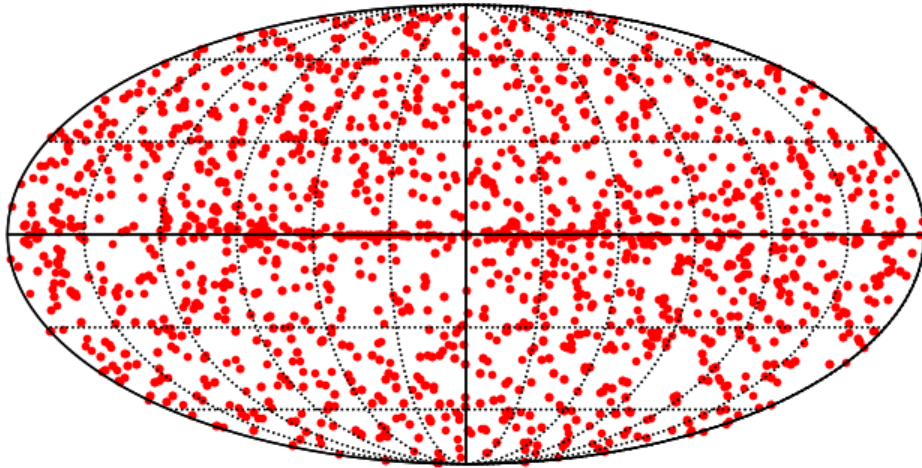


Figura 24: Ubicación de la fuentes compactas detectadas sobre las imágenes publicadas en 2015 por Planck del LFI tomadas a 44 GHz. Cada línea representa un intervalo de 30° .

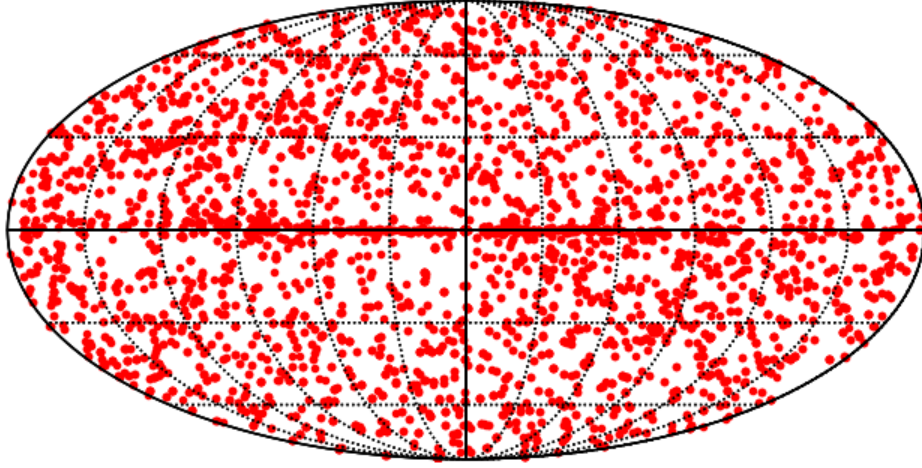


Figura 25: Ubicación de la fuentes compactas detectadas sobre las imágenes publicadas en 2015 por Planck del LFI tomadas a 70 GHz. Cada línea representa un intervalo de 30°.

Comparamos ahora los flujos proporcionados por el PCCS2 y nuestros catálogos para las fuentes que aparecen en ambos catálogos. En este caso hacemos un ajuste lineal de la forma:

$$y = ax + b, \quad (20)$$

donde a es la pendiente, b la ordenada en el origen, y es el flujo de las fuentes detectadas por nosotros, y x es el flujo de las mismas fuentes pero aportado por Planck. a nos informa de como se parecen nuestra estimación de flujos con respecto a la de Planck, cuanto más cercano a la unidad más se parecen. b nos informa por otra parte de un error sistemático entre nuestras fuentes, una ordenada en el origen de +1 Jy nos indicaría por ejemplo que estamos realizando una sobre estimación de 1 Jy en nuestro catálogo. Conociendo a y b y usando la ecuación 20 podemos estimar la intensidad que mediríamos con nuestro filtro de una fuente del PCCS2 de intensidad conocida. Hacemos dos ajustes, uno imponiendo una ordenada en el origen nula y otro dejando b como parámetro del ajuste, añadiendo un grado más de libertad. Obtenemos lo siguiente:

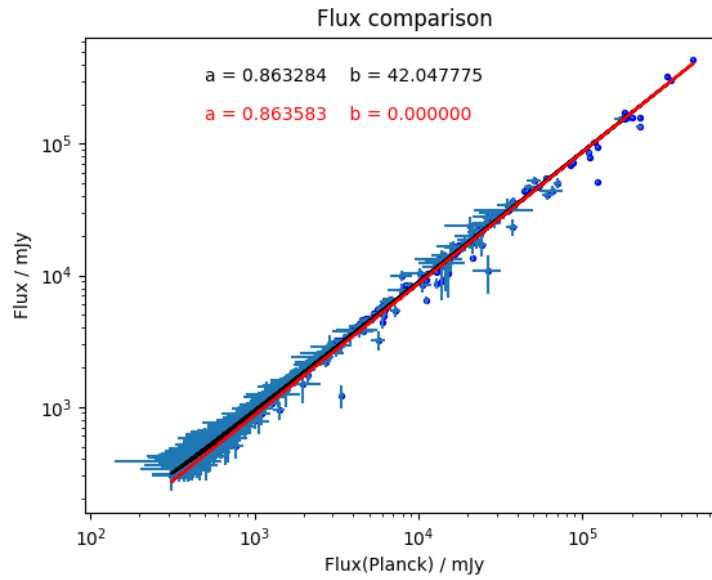


Figura 26: Ajuste lineal a las relaciones entre flujos obtenidos para las misma fuentes por nuestro código y por el catálogo de Planck, PCCS2, a 30 GHz.

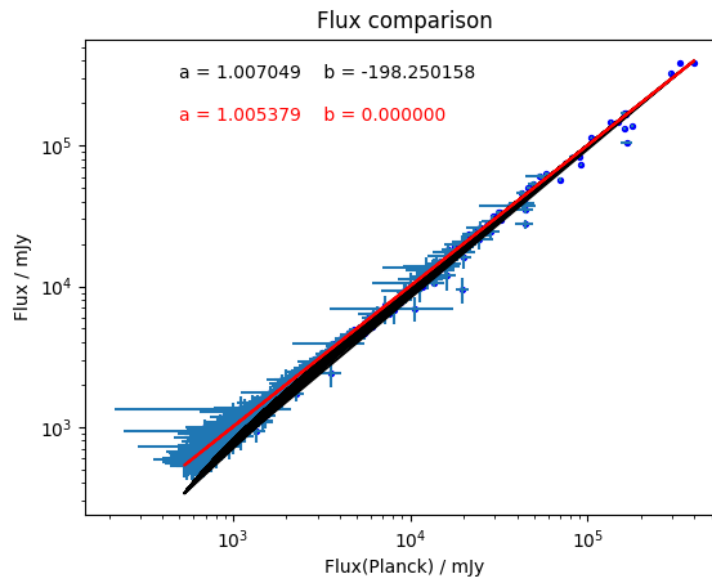


Figura 27: Ajuste lineal a las relaciones entre flujos obtenidos para las misma fuentes por nuestro código y por el catálogo de Planck, PCCS2, a 44 GHz.

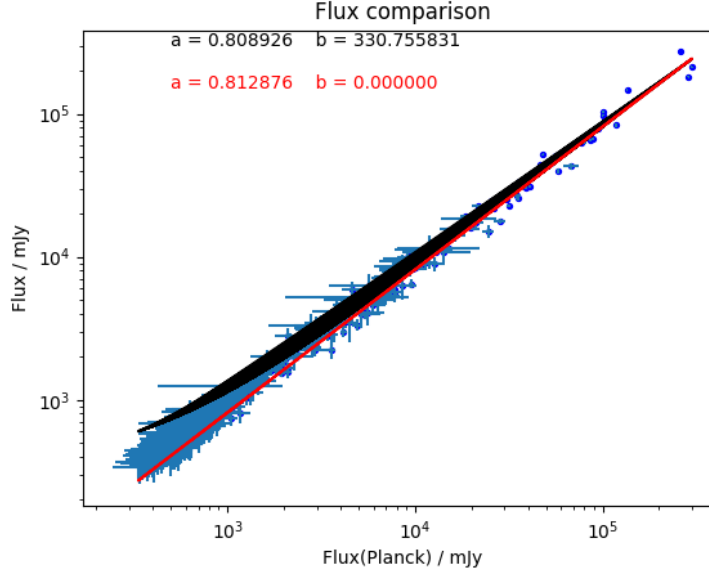


Figura 28: Ajuste lineal a las relaciones entre flujos obtenidos para las mismas fuentes por nuestro código y por el catálogo de Planck, PCCS2, a 70 GHz.

Sabemos que a debe parecerse mucho a la unidad dado que los datos del PCCS2 los tomamos por válido y buscamos que los nuestros sean lo más parecidos posible. Podemos observar que, salvo el caso de 44 GHz, las pendientes distan mucho de parecerse a la unidad tomando valores como 0,8. Esto señala que algo no parece ir bien entre nuestra forma de obtener los flujos y la de Planck. Por otro lado vemos que en los casos de 30 y 44 GHz estamos sobrestimando y subestimando el flujo respectivamente, no mucho unos 40 y 200 mJy respectivamente (dentro de los márgenes de error). Hay que destacar que aunque los errores parezcan mucho mayores para flujos bajos (los errores relativos son más mayores), no es tan grande como se pueda pensar observando las figuras 26, 27 y 28, ya que estas están en escala logarítmica.

5.1.1 Ajuste método RANSAC

Lo primero que podemos pensar es que quizás las fuentes que hemos detectado y que parecen coincidir con las del PCCS2 no son las mismas, por eso las discrepancias. Esto no parece ser el caso ya que observamos una relación lineal. Si estuviéramos comparando fuentes distintas sería muy improbable que los errores cometidos se relacionaran de una manera lineal. Este podría ser el caso de unas pocas fuentes las cuales si habríamos tomado mal. Queremos entonces un método que nos permita distinguir esas fuentes y eliminarlas, así podríamos comparar las restantes y ver si en este caso nuestras estimaciones del flujo son las correctas (comprobando así la validez de nuestro método).

Utilizamos para esto el algoritmo RANSAC, RANDOM Sample Consensus. Este método es un algoritmo iterativo que estima los parámetros de un modelo, en nuestra caso uno lineal, a partir de una serie de datos. En estos datos asumimos que hay “outliers” (datos malos) que no se ajustan al modelo y pueden ser considerados como inválidos o ruido que sobra en nuestro ajuste. Asumimos también que estos outliers no influyen en el resto de datos (Cierto, las fuentes están por lo general a distancias angulares lo suficientemente grandes como para no influir unas a otras). Podemos utilizar este método tanto para encontrar estos outliers, detección de datos incorrectos o ajenos a un proceso, como para ajustar el resto de datos a nuestro modelo. Es un algoritmo no determinista, da resultados con una probabilidad que aumentando con el número de iteraciones del método.

Una asunción que realizamos es que la gran cantidad de los datos son datos que se ajustan al modelo, nuestro caso cumple este requisito (siguen una relación lineal). RANSAC es un modelo iterativo. Selecciona un conjunto aleatorio entre nuestros datos. Los ajusta a nuestro modelo e incluye en el subconjunto los datos fuera de este que se ajustan lo suficiente (dado un criterio) al modelo. Si se incorporan suficientes datos (de nuevo, dado otro criterio) se considera

como una buena iteración. Se realiza de nuevo el ajuste con el nuevo subconjunto de datos. Se realiza de manera iterativa este conjunto de operaciones. Si en algún ajuste datos del subconjunto no se ajustan como deberían al modelo (otro criterio) se eliminan del conjunto. Si llegamos al número máximo de iteraciones (parámetro interno del método) o el subconjunto no varía lo suficiente entre iteraciones el método para. Filtra el subconjunto final asumiendo que los datos fuera de este son los outliers. Este método original de 1981 fue propuesto por Fischler y Bolles, ver [22].

Este algoritmo es tremendamente útil en análisis de imágenes y en el, relativamente nuevo, campo del "machine learning" que puedes entrenar a un algoritmo a reconocer datos que se ajusten a un modelo utilizando EL RANSAC. Afortunadamente dicho algoritmo ya está implementado en Python, al igual que en otros muchos lenguajes de programación. Podemos ver en la figura 29 un ejemplo del uso de este algoritmo.

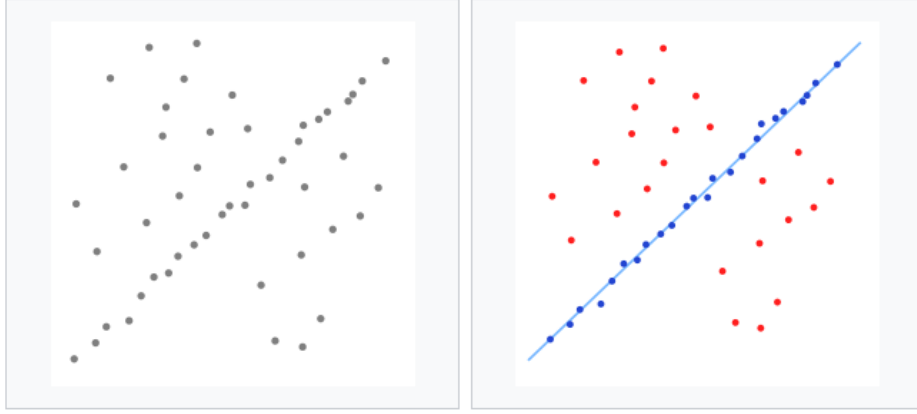


Figura 29: Conjunto de datos con outliers antes y después de ser ajustados linealmente por el algoritmo RANSAC. Los outliers no influyen en el ajuste. Fuente: <https://es.wikipedia.org/wiki/RANSAC>.

Utilizamos este método para comparar las coincidencias entre nuestros catálogos y los del PCCS2 en las frecuencias del LFI. Obtenemos lo siguiente:

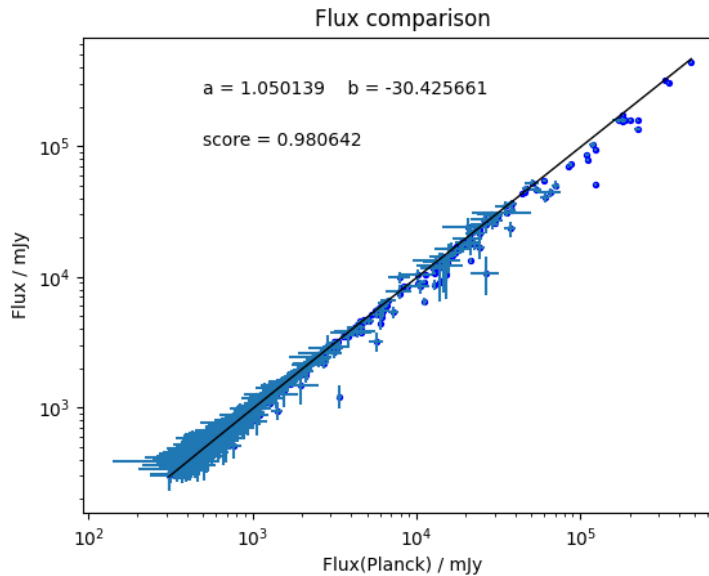


Figura 30: Comparación de flujos de las fuentes coincidentes en nuestro catálogo y el PCCS2 a 30 GHz mediante un ajuste lineal utilizando el método RANSAC.

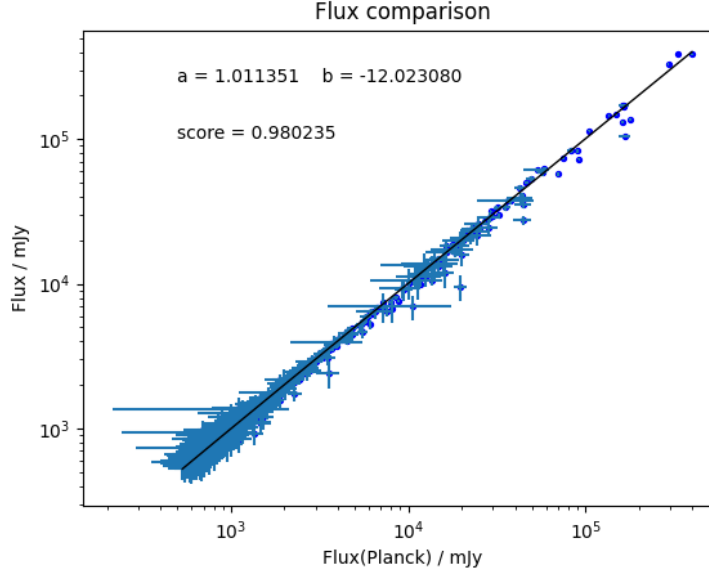


Figura 31: Comparación de flujos de las fuentes coincidentes en nuestro catálogo y el PCCS2 a 44 GHz mediante un ajuste lineal utilizando el método RANSAC.

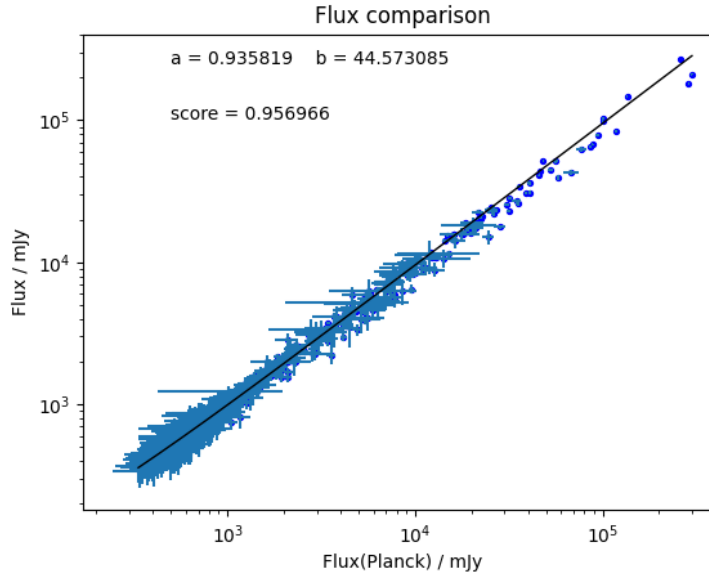


Figura 32: Comparación de flujos de las fuentes coincidentes en nuestro catálogo y el PCCS2 a 70 GHz mediante un ajuste lineal utilizando el método RANSAC.

El parámetro score, proporcionado por la rutina RANSAC implementada en Python, da una medida de la calidad del ajuste. Esto indica que tenemos unos errores aproximadamente de un 2% para los ajustes a las frecuencias de 30 y 44 GHz y uno de un 5% para el ajuste a la frecuencia de 70 GHz.

Tenemos ahora unos ajustes con pendientes más parecidas la unidad, coincidentes dentro del error. Nos indica que la estimación del filtro que realizamos es compatible con la que se realizó en Planck en su momento. Sin embargo, el tener que recurrir al método RANSAC nos produce un sabor agridulce. Como estamos eliminando datos para que el ajuste funcione estamos, en cierto modo, trampeando ya que esos datos que eliminamos y hemos detectado coinciden con

datos de Planck. Esto nos dice que algo falla en nuestro método de búsqueda o en el filtro en determinadas situaciones.

5.1.2 Ajuste método ODR

Antes de buscar los errores que sugieren nuestro último ajuste intentamos algo distinto. En nuestro ajuste original sin eliminar datos no hemos pesado nuestros datos. Tanto el PCCS2 como nuestro catálogo tienen una columna con una estimación del error de los flujos. Buscamos ahora un método para realizar un ajuste lineal en el cual se pesen los errores. Al ser la mayor cantidad de fuentes compatibles en ambos catálogos fuentes débiles con una peor estimación del flujo, mayor incertidumbre, puede que nos este empeorando el ajuste.

Utilizamos el método ODR, Orthogonal Distance Regression. Usamos este método por que tenemos errores tanto en la variable x como en la y , es decir, en la variable dependiente e independiente. El problema es que el método de mínimos cuadrados con los que generamos nuestras rectas de ajuste (busca el mínimo error cuadrático medio, es decir, la media de las distancias al cuadrado entre nuestros puntos y sus valores esperados en la recta del ajuste), trata a las variables independientes como fijas. En este caso al tener una incertidumbre tanto en las variables dependiente e independientes el método de mínimos cuadrados ordinario no funciona de manera óptima.

ODR es un algoritmo que busca los mejores parámetros de ajuste a un modelo teniendo en cuenta los errores, los pesa en cierto modo. Busca minimizar el error cuadrático medio pero ligeramente cambiado. No minimiza la suma de las distancias (de un punto a la recta del ajuste) al cuadrado, minimiza la suma de las distancias al cuadrado mínimas teniendo en cuenta todos los posibles valores de cada variable. Cuando conocemos los valores de ambos errores el método ODR es mejor que el de mínimos cuadrados.

El ODR es un problema relativamente complejo, pero está implementado en Python y es fácil de usar. Para saber más del problema ODR y de su comparación con el método de mínimos cuadrados, ver [23]. Los ajustes obtenidos mediante el algoritmo ODR fueron:

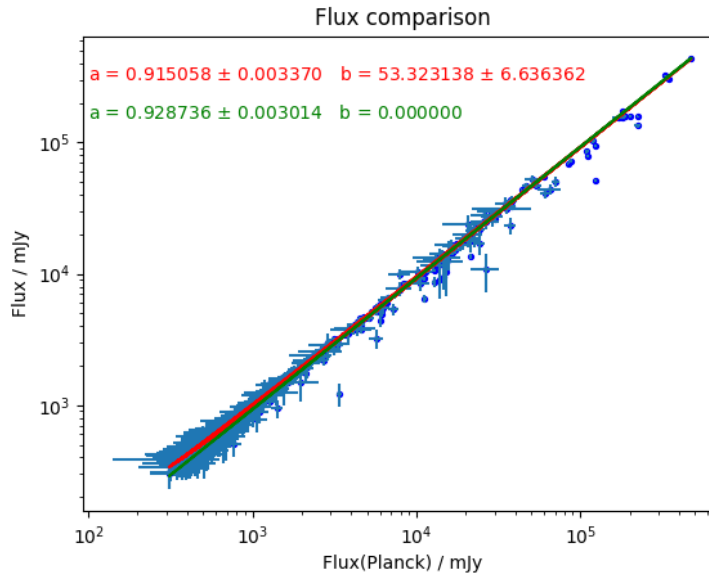


Figura 33: Comparación de flujos de las fuentes coincidentes en nuestro catálogo y el PCCS2 a 30 GHz mediante un ajuste lineal utilizando el método ODR.

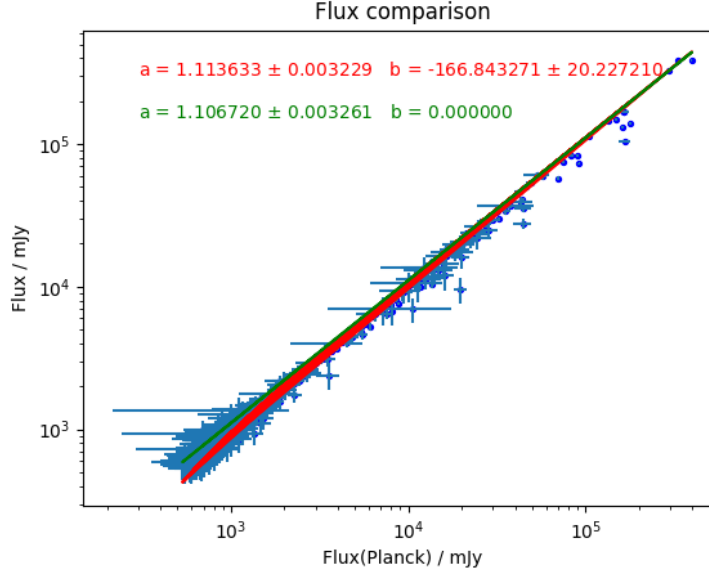


Figura 34: Comparación de flujos de las fuentes coincidentes en nuestro catálogo y el PCCS2 a 44 GHz mediante un ajuste lineal utilizando el método ODR.

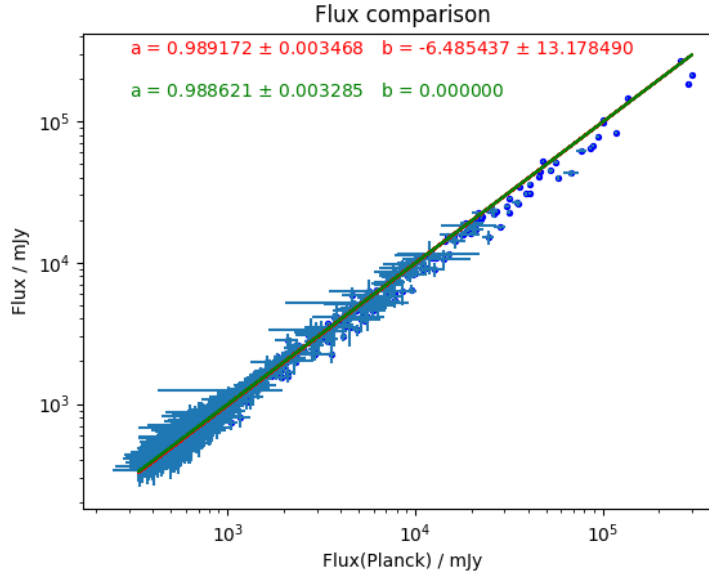


Figura 35: Comparación de flujos de las fuentes coincidentes en nuestro catálogo y el PCCS2 a 70 GHz mediante un ajuste lineal utilizando el método ODR.

El ajuste mediante el método ODR parece funcionar. No se ajusta tanto como el RANSAC, aunque en este caso no estamos eliminando datos y obtenemos pendientes entre 0,9 y 1,1. Tenemos pendientes relativamente cercanas a la unidad, más que con el ajuste lineal por mínimos cuadrados. Sabemos también que al estimar el flujo de forma distinta es normal que las pendientes no sean iguales a uno (Planck usó un filtro Mexican-Hat Wavelet 2, nosotros un filtro adaptado. Por otra parte la forma de distribución de los parches es distinta en ambos casos). Por todo esto nos quedamos con el ajuste del ODR, el cual demuestra que nuestro método de detección funciona bien y queda corroborado por los catálogos del Planck PCCS2 del LFI.

5.2 Estudio de fuentes mono-catálogo

Un análisis mayor de nuestros resultados requiere que analicemos las fuentes que detectamos en nuestros catálogos y que Planck no pudo detectar y viceversa. Este análisis lo llevamos a cabo sobre los catálogos realizados con un corte de detección igual a Planck, un SNR igual a 4. Con el objetivo de resumir expondremos solamente el caso de 30 GHz, siendo los otros dos análogos.

Nuestro catálogo contiene 1378 fuentes, el PCCS2 a 30 GHz contiene 1560 (parecen ser similares en tamaño). Podemos explicar que el PCCS2 tenga una extensión ligeramente mayor aun usando el mismo umbral de detección. Para generar estos catálogos antes de utilizar su método (en el cual se basa este trabajo) realizaron varios filtrados más, sobre todo en el plano galáctico, con el objetivo de maximizar el SNR en esa zona. Esto permitió detectar un número de fuentes ligeramente superior. Por otro lado las ya mencionadas fluctuaciones del ruido que varían ligeramente al filtrar y sobre todo al proyectar parches en zonas distintas. Pueden hacer que una fuente débil que nosotros detectamos justo por encima del umbral de detección Planck no la viera y viceversa. También fuentes muy brillantes pueden llegar a estropear el filtro (saturando la imagen) haciendo que no consigamos ver alguna que vio Planck o el caso contrario. Esto solo ocurriría con fuentes muy brillantes y sólo con unas pocas, si viéramos esto en grandes cantidades algo estaría mal. Finalmente los PCCS2 también cuentan con filtrados realizados para beams no gaussianos, ligeramente deformados. Además estos catálogos incluyen unas pocas fuentes algo más extensas, sin dejar de ser fuentes compactas (algo en el límite). El PCCS2 incluye una columna indicando de que tipo es la fuente.

Nuestra hipótesis es que Planck encuentra más fuentes debido a esto último. Para demostrar nuestra hipótesis deberíamos observar que las fuentes que Planck detecta y nosotros no recaen principalmente en el plano galáctico (debido a esos filtrados extra). Debería verse que las fuentes que aparecen sólo en un catálogo son en su mayoría débiles, con alguna excepción de fuentes muy brillantes (Notar que en este contexto con brillo nos referimos a como destaca en su entorno, su SNR, no a su emisión absoluta en Jy. Una fuente de 2 Jy puede ser muy brillante en un entorno con poca emisión, e imperceptible en el centro de nuestra galaxia).

Empezamos dibujando las fuentes que aparecen únicamente en un catálogo:

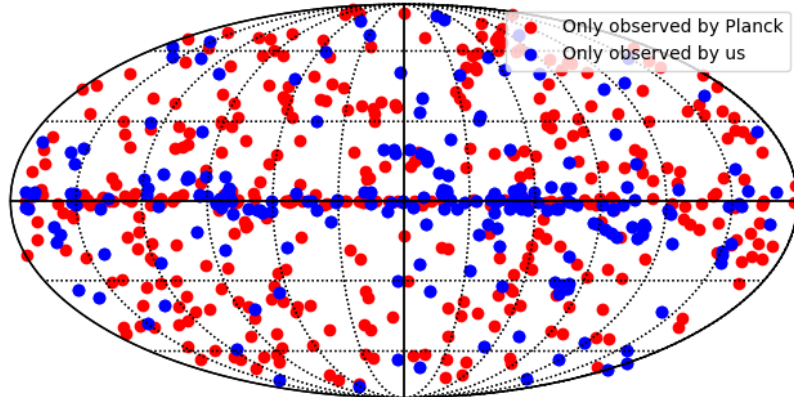


Figura 36: Localización de las fuentes mono-catálogo en la esfera celeste. Cada línea representa una división de 30° .

Podemos observar que la mayoría de fuentes detectadas solo por un catálogo caen en el plano galáctico, lo cual concuerda con nuestra hipótesis. Allí la alta emisión de la galaxia hace que el SNR sea, por lo general, menor además de ser la zona con filtrados extra por Planck.

Para acabar de confirmar, o desmontar, nuestra hipótesis queda realizar un histograma de las fuentes mono- catálogo.

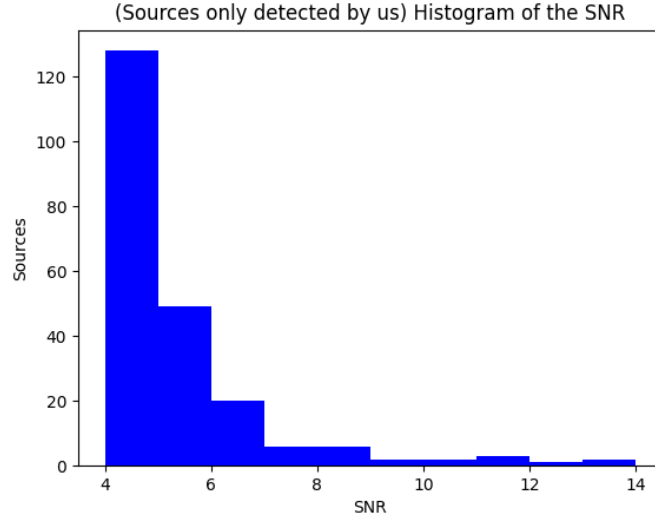


Figura 37: Histograma de la señal ruido, SNR, de las fuentes detectadas únicamente por nosotros.

Observamos que la gran cantidad de fuentes detectadas solo por nosotros son fuentes débiles, justo por encima del umbral de corte, $\text{SNR} = 4$. Otras fuentes, muy pocas, tienen SNR superiores a 10 (relativamente brillantes). Finalmente unas pocas caen en rangos intermedios, quizás debido a las sutiles diferencias entre los filtros. La media del SNR es de $5,0 \pm 1,6$.

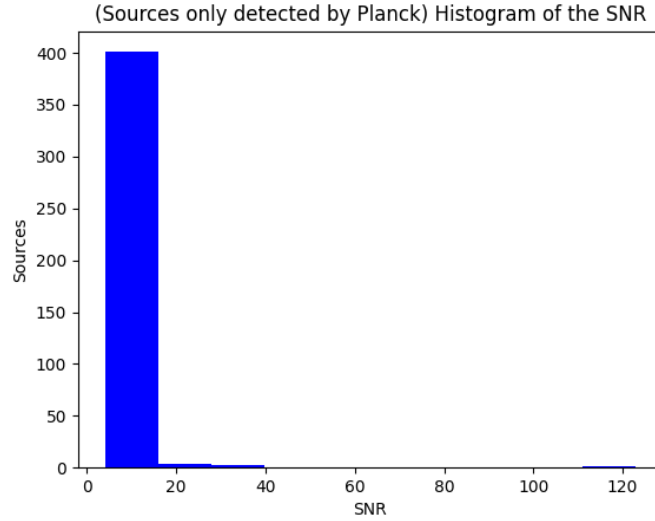


Figura 38: Histograma de la señal ruido, SNR, de las fuentes detectadas únicamente por Planck.

Nuevamente la mayor cantidad de las fuentes exclusivas de este catálogo tienen valores señal ruido justo por encima del umbral de detección, a excepción de unas pocas extremadamente brillantes, $\text{SNR} > 10$. La media del SNR fue en este caso de $5,8 \pm 6,5$, la cual esperábamos ligeramente mayor al poder detectar fuentes en el plano galáctico (debido al limpiado extra en esa zona) de valores de SNR en un amplio rango. Resta mencionar que una combinación de ambos catálogos da un valor completo, ya que lejos de sustituirse el uno al otro, se complementan.

Llegados a este punto hemos conseguido nuestro objetivo: un método para detección de fuentes compactas de origen extragaláctico, implementado en Python, y que nos ofrece catálogos más extensos aunque que menos detallados que los del PCCS2. Sirven para complementar a estos. Principalmente el éxito de obtener catálogos más extensos se debe

a que este filtro permite bajar ligeramente el umbral de detección. También se puede coger el mismo que Planck para estar más seguros de no tener alguna falsa detección. En ese caso seguimos con un catálogo de extensión ligeramente menor pero que sigue complementando al PCCS2. Este método es funcional y puede utilizarse con futuras imágenes del FCM a distintas frecuencias y FWHM.

5.3 Número de cuentas en función del flujo

En este punto ya hemos conseguido el objetivo del trabajo. Aprovechando que lo hemos conseguido en un plazo de tiempo aceptable hemos decidido hacer un análisis sobre las fuentes compactas obtenidas con un umbral de detección de $3,5\sigma$ (SNR=3.5). Dicho análisis compara el número de cuentas diferenciales, el número de fuentes S en nuestros catálogos en intervalos de flujo ($\log_{10}(dN/d\log S)$ [sr^{-1}], frente a $\log_{10}S$), con el esperado según unos modelos teóricos.

De las fuentes compactas que detectamos conocemos muy poco sobre los procesos tan complejos que hacen que emitan en nuestras frecuencias. Por esto la necesidad de obtener catálogos más amplios. Ampliando la estadística facilita la mejora de nuestra información sobre dichas fuentes. No tenemos fórmulas analíticas que con las condiciones iniciales oportunas nos informen de como es la forma de dN/dS , número de fuentes por unidad de diferencial de flujo, para cada uno de estos objetos. Trabajamos entonces a la inversa, a partir de los datos generamos modelos semi-analíticos que intentan reproducir las observaciones. Con los datos suficientes podemos crear estos modelos empíricos que se ajustan muy bien a la realidad en el rango en el que se crearon. Esto permite hacer predicciones sobre la cantidad de fuentes que encontraremos en estos rangos de intensidad a una frecuencia en una determinada zona del cielo. Esto no es nada nuevo ni novedoso de la astrofísica, de hecho en casi todos los campos de la física, si no en todos, se trabaja en algún momento de esta manera. A partir de los datos sacamos conclusiones, creamos modelos empíricos y obtenemos la capacidad de predecir en estos rangos. Si bien es cierto que buscamos la base teórica que hay detrás para explicar nuestros datos y ampliar el espectro de predicciones, en muchos casos necesitamos de la observación en un primer momento y de modelos empíricos para dar el primer paso hacia las ideas subyacentes.

En 2005 de Zotti et al. [24] generaron a partir de los datos disponibles en ese momento los modelos para cuentas diferenciales (y totales). Son válidos hoy en día hasta frecuencias de hasta 100 GHz. Tucci et al. [25] extendieron estos modelos hasta frecuencias de hasta unos 900 GHz para unas frecuencias determinadas. Dichos modelos para las frecuencias del LFI toman el aspecto que aparece en la figura 39.

Estos modelos consideran por separado las galaxias por tipos. Por ejemplo, a bajas frecuencias distinguen entre radiofuentes de espectro plano (flat-spectrum) y espectro no plano (steep-spectrum). A frecuencias más altas aparecen también otro tipo de fuentes como fuentes de espectro invertido, blazars (BL-Lac) y flat-spectrum radio quasars (FSRQ), ver [26]. Los modelos toman todos esos tipos de fuentes por separado y calculan el número total sumando todos los tipos.

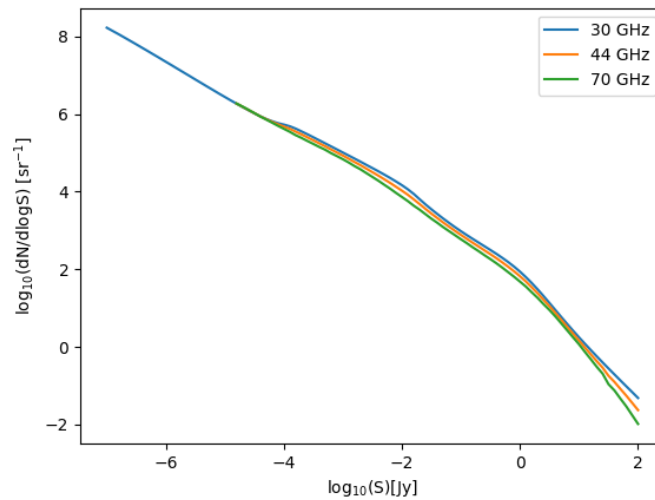


Figura 39: Modelos teóricos de cuentas diferenciales para las frecuencias del LFI de Planck.

En los años anteriores a la misión Planck no se conocía apenas nada sobre como se comportaban estas fuentes. Hoy en día los modelos entre 1 y 100 GHz con flujos no inferiores al milijansky están altamente validados y encajan perfectamente con los datos obtenidos. Afortunadamente estos modelos cubren las frecuencias del LFI por lo que podemos comparar nuestros datos con los modelos. Esto sirve como verificación extra además de la previa comparación con el PCCS2. Podemos también testear nuevamente el modelo con nuevos datos afianzándolo aun más si cabe.

Los modelos de de Zotti, Tucci et al.[24],[25] contabilizan fuentes extragalácticas únicamente. Sin embargo en nuestro catálogo habremos detectado, sin poder evitarlo, objetos galácticos (nebulosas, remanentes de supernova, etc) que se añadirán a los números de objetos que encontremos. Esto se notará mucho más en las zonas cercanas al plano galáctico (donde se condensa la mayor parte de objetos galácticos) que en latitudes altas.

La presencia de fuentes galácticas tiene a contaminar los números de cuentas haciendo que parezcan mayores de lo que son. Conviene en nuestros análisis, por seguridad, usar solamente las fuentes a latitudes no muy próximas a cero, evitando así fuentes galácticas.

Las cuentas diferenciales vienen dadas por estereorradián, es decir, por unidad de ángulo sólido. Nuestro un catálogo cubre todo el cielo, un área de 4π estereorradianes. Podríamos pensar que para comparar con los modelos teóricos deberíamos dividir el numero de cuentas por este área, pero no es así. Nuestro catálogo cubre toda la esfera de forma no homogénea. a bajas latitudes, cerca del plano galáctico, los ruidos de nuestra imagen (instrumental + CMB + galaxia) son mayores. Para un mismo umbral de detección (SNR dado) encontraremos pocas fuentes débiles, mientras que a latitudes muy altas estaremos obteniendo muchas de las fuentes débiles que no observábamos en el plano galáctico. Sin filtrar en mayor medida esos ruidos extra, que no penalizan tanto en latitudes altas, no podemos eliminar este sesgo. Por todo esto nuestro número de fuentes extragalácticas a un determinado flujo puede que sea más pequeño de las que realmente hay.

La aparición de este sesgo hace que el número de cuentas, al menos para flujos bajos, parezca menor de lo que es. Esto podría corregirse calculando un área efectiva menor que 4π que corrigiera las cuentas. Dicha corrección sería costosa y llevaría un tiempo extra. Hemos decidido no hacerla.

Para reducir estos dos sesgos (fuentes galácticas, no homogeneidad de cuentas) evitamos por completo la zona galáctica. Cortamos y dejamos fuera del catálogo todo lo que tenga latitud galáctica $|b| < b_c$, donde b_c es un determinado corte que asignamos. El área que dejamos cuando hacemos tal corte es

$$A = 4\pi \int_0^{\pi/2 - b_c} \sin \theta d\theta = 4\pi(1 - \sin b_c) \quad (21)$$

El ángulo b_c va expresado en radianes. Si no pongo ningún corte, $b_c=0$, recupero $A=4\pi$, si impongo un corte $b_c = 30^\circ$, $\pi/6$ rad, obtengo $A=2\pi$, es decir, la mitad del cielo. Se debe hacer un compromiso, entre elegir una zona muy pequeña y muy limpia o una gran extensión pero altamente contaminada. Escogemos un ángulo de corte b_c de 30° , la mitad de la esfera celeste, ver figura 40. Lo cierto es que la contaminación galáctica no sigue exactamente una distribución por bandas regulares de latitud. Hay regiones contaminadas a latitudes altas, por ejemplo las nubes de Magallanes o el llamado Galactic Spur. De todas formas vamos a realizar este análisis aunque, con tiempo, podría realizarse uno más fino incluyendo más fuentes y mejorando la precisión.

Comparamos con los modelos teóricos:

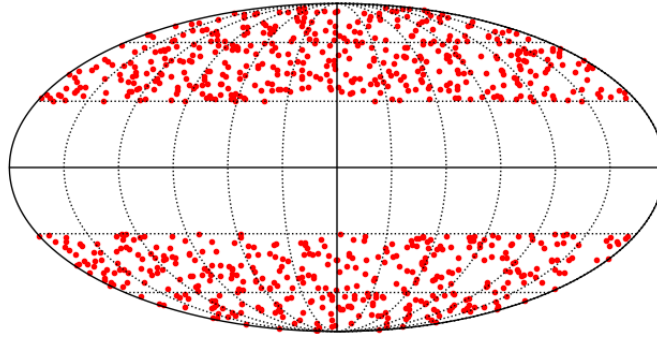


Figura 40: Fuentes compactas en nuestro catálogo con latitudes galácticas superiores a 30° . Frecuencia del catálogo 30 GHz.

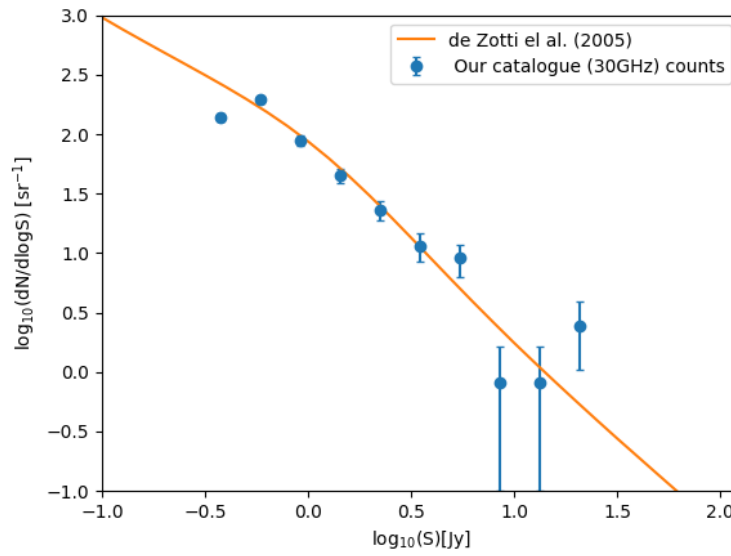


Figura 41: Comparación de las cuentas diferenciales obtenidas de nuestro catálogo de 30 GHz ($\text{SNR} \geq 4$) con los modelos teóricos.

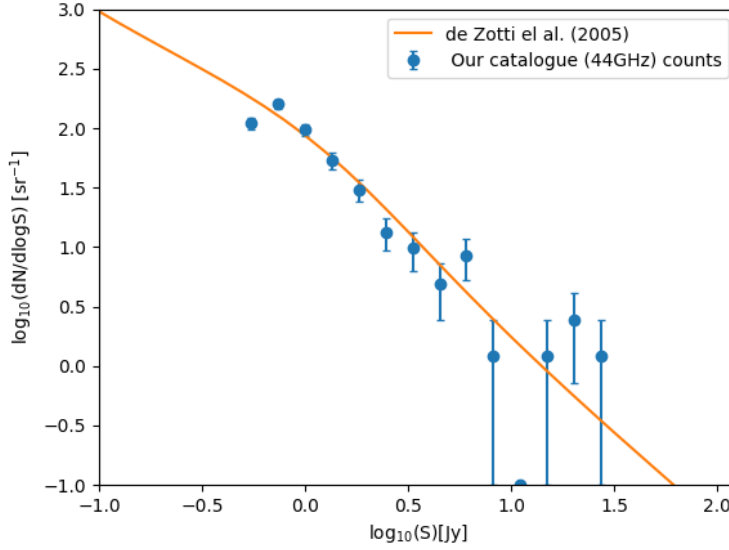


Figura 42: Comparación de las cuentas diferenciales obtenidas de nuestro catálogo de 44 GHz ($\text{SNR} \geq 4$) con los modelos teóricos.

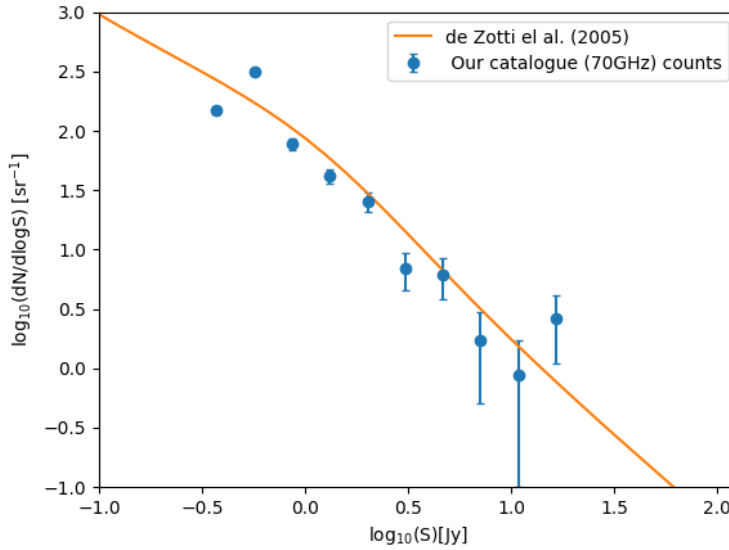


Figura 43: Comparación de las cuentas diferenciales obtenidas de nuestro catálogo de 70 GHz ($\text{SNR} \geq 4$) con los modelos teóricos.

Las rutinas para calcular las cuentas diferenciales a partir de un catálogo fueron proporcionadas por el director de este TFG.

Distingamos tres regiones:

- En los flujos más altos, en este caso en torno a $\log_{10} S \sim 1$ ($S \sim 10$ Jy), las cuentas se elevan ligeramente por encima de la curva teórica. Probablemente se deba a que no hemos eliminado todas las fuentes galácticas ni aquella provenientes de zonas como las nubes de Magallanes o el Galactic Spur. Esto aumenta las cuentas en esos rangos

de intensidad. Por otra parte tenemos pocas fuentes con estas intensidades, es normal que no se ajuste bien al modelo.

- En flujos intermedios ($-0.4 \leq \log_{10} S \leq 1$), los puntos se ajustan bien a la curva. Esto sugiere que nuestro catálogo está recogiendo bien este tipo de fuentes en esta región del cielo.
- Por debajo de $\log_{10} S = -0.4$ el número de cuentas del catálogo cae por debajo de la curva. Esto puede deberse a que estamos cerca del límite del método de detección y empezamos a dejar de ver las fuentes que hubiera en esos flujos. El límite de completitud del catálogo, flujo por encima del cual recogemos todas las fuentes del catálogo, es aproximadamente $S_{comp} \simeq 10^{-0.4} \sim 400 \text{ mJy}$.

Veamos la comparación de unos de los PCCS2, en este caso el de 30 GHz, con los modelos teóricos:

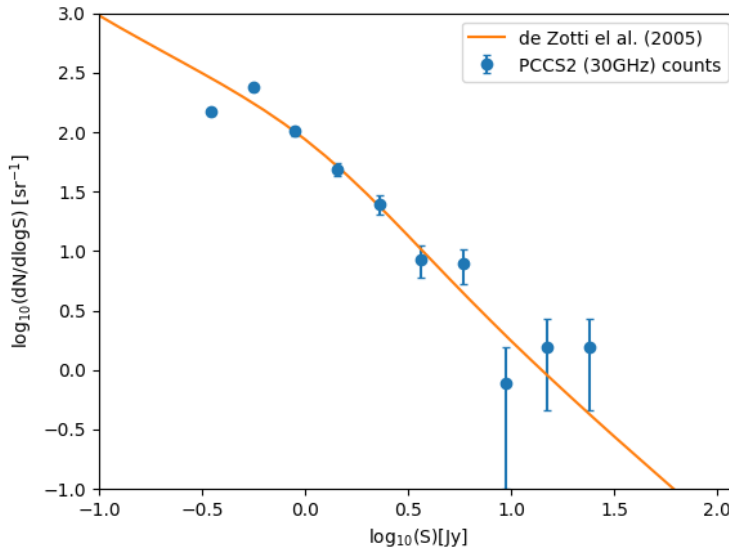


Figura 44: Comparación de las cuentas diferenciales obtenidas del PCCS2 30 GHz con los modelos teóricos.

Vemos que se parece mucho a la obtenida por nosotros. Esto corrobora una vez más la eficiencia de nuestro método. Comprobamos que el límite de completitud es similar, unos 400 mJy, lo que sugiere que nuestro filtro al permitirnos bajar el umbral de detección no nos hace ver fuentes más débiles, si no que nos permite ver fuentes débiles en zonas más ruidosas.

Los errores en las cuentas diferenciales, provenientes de la rutina aportada por el director del TFG, se calculan suponiendo que el número de cuentas que caen en un determinado intervalo de flujo son la realización de un proceso aleatorio de distribución de Poisson.

Se ha comprobado mediante un exhaustivo análisis de los datos que el método funciona y da catálogos compatibles con los de Planck aumentando ligeramente su extensión. La comparación de los catálogos así como la de las cuentas diferenciales con los modelos teóricos fueron desarrolladas por la misión Planck en su Intermediate Results [27].

6 Tiempos de ejecución. Comparación con método en Fortran95.

La razón de realizar esta implementación en Python se debe principalmente al uso de este lenguaje por la comunidad. Lo cierto es que Python es un lenguaje de alto nivel, esto quiere decir que es un lenguaje fácil de comprender (de sintaxis sencilla). Python es más intuitivo pues su sintaxis es más parecida al lenguaje humano lo que permite hacer códigos más

elegantes amistosos frente a primeras lecturas (es más “user firendly”). Todo esto no viene con un coste, la eficiencia. Los lenguajes de alto nivel son interpretados. Esto quiere decir que es código se traduce primero a un pseudo lenguaje máquina y después a lenguaje máquina de forma que nuestro ordenador lo entienda. En los lenguajes de bajo nivel como Fortran o C, pese a ser más complicados, el código se traduce directamente a código máquina. Esto hace que sean mucho más rápidos. Se puede hablar más de las diferencias entre alto y bajo nivel pero la diferencia principal que nos atañe es la velocidad.

Cambiamos a Python, pese a ser más lento, por que es un lenguaje en auge, en constante desarrollo, y más inclusivo. Python es uno de los lenguajes favoritos por la comunidad científica. Una ventaja es que la comunidad ha creado rutinas y funciones base de Python implementadas en C (lenguaje de alto nivel) con lo que son más rápidas de lo que esperaríamos.

Otro motivo del uso en Python es que permite hacer todo lo que necesitamos para el análisis de datos: filtro, figuras, o análisis de datos. El método antiguo, pese a estar implementado en Fortran, recurría a lenguajes para realizar tareas específicas como dibujar los mapas.

El tiempo de ejecución de este método, en el ordenador del alumno, a rondado entorno a los 7 y 10 minutos, dependiendo del mapa filtrado. los tiempo de la versión anterior rondaban los 5 minutos. No nos sorprende que tarde más. El incremento no es mucho por lo que estamos contentos.

Aproximadamente la mitad del tiempo de ejecución se consume en leer el mapa, parchear, y filtrar cada parche (4 a 5 minutos). La otra mitad se gasta en la eliminación de recurrencias en nuestro catálogo inicial. Al aumentar la resolución, aumentando la frecuencia por ejemplo, aumenta la cantidad de fuentes detectadas y las recurrencias en nuestro catálogo. Por ello el tiempo que nos consume eliminarlas aumenta(2 a 5 minutos).

El método para eliminar recurrencias ideado por el alumno permite eliminar las recurrencias en una cantidad reducida de tiempo, simplificando el código y dando una mejor estimación de los flujos, hecho a destacar de esta nueva versión.

Cabe esperar que al utilizar este código en futuras imágenes de más alta resolución los tiempos de ejecución aumenten. Por otra parte con los años la obtención de ordenadores más potentes contrarrestarán dicho aumento.

7 Fases del proyecto

En este trabajo se pueden diferenciar varias fases con diferentes objetivos. Es la suma de estas partes que ha permitido el correcto desarrollo de este trabajo.

Fase 1

Esta fase fue la más extensa en el tiempo, aunque no la que más horas a requerido. Se ha compuesto por una introducción del alumno en el ámbito en el que se desarrolla este trabajo.

Se pueden distinguir dos líneas de trabajo diferentes en esta fase:

- En primer lugar se ha realizado una lectura de distintos papers en los que se discuten: el tipo de fuentes que encontramos en los dominios de las ondas de radio de alta frecuencia y las microondas (frecuencias en las que hemos trabajado); la utilidad de los mapas del FCM para detectar fuentes compactas; los problemas de la detección de fuentes en estas frecuencias donde los límites de resolución de los que disponemos son del orden del tamaño de las fluctuaciones del FCM.
- Por otro lado se ha ido introduciendo al alumno en el lenguaje de programación de Python. Si bien es cierto que debido al grado de física de la UC el alumno sabía programar, este solo conocía los lenguajes de java y Matlab (interesantes e importantes en distintos ámbitos). Sin embargo de Python no se tenía conocimiento alguno. El aprendizaje fue rápido y sencillo.

Fase 2

En esta segunda fase el director del trabajo fue proporcionando las herramientas necesarias para el código. Entre estas herramientas se incluyen el filtro o las librerías de Healpy, Astropy y Scikit entre otras.

El método de trabajo en esta fase ha sido siempre del siguiente estilo. El director ha ido proporcionando la herramienta con algún tipo de ejemplo y unos apuntes explicando brevemente el funcionamiento. Con todo esto el alumno se ha leído las explicaciones, ejemplos incluidos, y documentaciones de las herramientas proporcionadas por los desarrolladores. Ha tratado de replicar los ejemplos con los mismo datos y datos distintos con el fin de familiarizarse con su uso. Finalmente el director le ha pedido algún tipo de tarea que ha realizado, como la optimización del parámetro NBIN en el filtro.

Esta fase ha sido, sin lugar a duda, la fase que más horas de trabajo a requerido.

Fase 3

Esta fase, la más importante quizás, ha tratado de la escritura del código. Una vez conocidos todas las herramientas el alumno se ha encargado de escribir cuidadosamente el código. Mediante prueba y error hemos conseguido un código funcional que nos da como resultado unos datos.

Esta fase pese a no haber requerido tanto tiempo ha sido quizás la más ardua. El alumno ha tenido que hacer uso de todo su ingenio y de las herramientas utilizadas para, por su cuenta, escribir el código. Antes de empezar a programar se ha discutido con el director la estrategia a tomar, valorando opciones. El alumno a presentado su idea de trabajo. El director ha sugerido algún cambio, como la búsqueda de coincidencias entre catálogos, y ha aprobado la estrategia. Es en esta fase el alumno ha aportado nuevas ideas como la resolución coincidencias.

Fase 4

Esta última fase recoge todos los análisis de datos. Se han detectando errores en el código y se han solventado.

Se ha comparado los catálogos obtenidos con los PCCS2 y las cuentas diferenciales con los modelos teóricos. Se detectaron dos errores principales: la necesidad del uso del RIMO (las frecuencias y FWHM efectivas), y un pequeño error en la rutina de cambio de píxel a coordenada.

Una vez terminadas todas estas fases el trabajo ha sido completado.

8 Hitos y conclusiones

En esta sección vamos a repasar los hitos del proyecto, comentar conclusiones sobre la implementación realizada y los datos obtenidos, comentar las habilidades obtenidas y hablar sobre las posibles extensiones del trabajo así como sus posibles usos futuros.

Hemos implementado en Python un método de detección de fuentes compactas en imágenes del FCM. Este método basado en una versión en Fortran ha probado un filtro diferente (filtro adaptado) al MHW2 que utilizó la colaboración Planck. Por motivos de eficiencia computacional (cálculo de la FFT) como de estadísticas (emisión de fonfo homogénea) se ha dividido la esfera en parches cuadrados (por proyección gnomónica) sobre los cuales se ha realizado el proceso de filtrado. Es imposible dividir la esfera en parches cuadrados de forma exacta. Por esto y para no perder fuentes por dejar zonas sin filtrar hemos decidido solapar parches (parches equidistantes). Esto ha traído consigo la repetición de fuentes en nuestros catálogos. Para la eliminación de repeticiones se ha implementado un nuevo método que parchea entorno a la zona de repetición y resuelve los conflictos. Hemos comparado nuestros catálogos con los de Planck y con modelos teóricos, mostrando un alto grado de compatibilidad. Pese a cambiar a un lenguaje interpretado y de más alto nivel que Fortran (Python) los tiempos de ejecución se han mantenido estables (de 5 a 7-10 min). Los catálogos que hemos producido fueron del orden de los de Planck para un mismo corte ($\text{SNR} \geq 4$) y más extensos reduciendo el

umbral ($\text{SNR} \geq 3, 5$), aunque quedaría comprobar si tenemos demasiadas fuentes espúreas. Este método pese a que ha sido probado solo con las imágenes de Planck del LFI (fase de ensayo. Escogemos estas frecuencias por comodidad), es válido para su uso con otros mapas del FCM tomados a distintas frecuencias y con otras FWHM. Un aspecto negativo es que pese a que al cambiar a Python buscábamos poder usar este método en cualquier sistema operativo, al usar Healpy sólo podemos usarlo en Mac IOS y Linux (no está implementado en Windows).

Afirmamos que el método funciona. Ha sido contrastado con datos y modelos exhaustivamente comprobados. Podemos decir con alegría que el proyecto ha sido un éxito, logrando las metas propuestas y compensando el esfuerzo y tiempo invertido.

Este trabajo ha servido al alumno para comprender más sobre el FCM y su importancia en la cosmología. Conocer las fuentes compactas que detectamos y su necesidad para comprender procesos como las primeras etapas en la formación de las galaxias. Por otro lado ha aprendido a programar en un lenguaje de programación cada vez más utilizada y exitoso, en física y otros mercados laborales. El alumno ha aprendido a tratar con datos, analizarlos, ubicar fallos en sus métodos, plantear hipótesis, y resolver problemas. Se ha familiarizado con algunas las herramientas más utilizadas en programación en los ámbitos de Astrofísica y cosmología. Ha descubierto algunos repositorios y bases de datos utilizadas en este ámbito. Ha aprendido como es parte del trabajo en estos campos y como funcionan macro-colaboraciones como Planck. Todo esto le ayuda prepararse para la iniciación en una vida laboral dedicada a la investigación. También le aporta cualidades tremendamente valiosas en otros nichos laborales.

Para finalizar este apartado comentemos las posibles expansiones de este proyecto, sus aplicaciones, y posible relevancia en otros trabajos.

Si bien es cierto que el método funciona, podría afinarse más (estamos en la versión 1.0, ahora entra en fase de extensión y mejora). Podrían aplicarse filtrados previos sobre las imágenes maximizando el SNR y permitiendo detectar más fuentes. Podrían aplicarse las correcciones del beam para filtrar. Se podrían realizar barridos a varias frecuencias contiguas mejorar la calidad de las detecciones. Una aproximación más ambiciosa podría realizarse aplicando métodos de machine learning. Crearía un algoritmo capaz de clasificar el tipo de fuente compacta dándole mayor profundidad a nuestros catálogos. Los métodos de parcheado y resolución de coincidencias podrían programarse en paralelo, reduciendo drásticamente el tiempo de ejecución. Estos y varios cambios más, sutiles o no tan sutiles, podrían realizarse con el objetivo de mejorar el método. La razón de que estos cambios no se hayan realizado es principalmente debido a que, o bien se escapan del nivel de dificultad de un trabajo de fin de carrera en física, o no se dispone del tiempo necesario para su implementación en un proyecto de 18 créditos ECTS.

Pese a que este proyecto se ha pensado para su uso sobre imágenes del FCM es cierto que puede ser utilizados para detección de fuentes en imágenes en las que se cumpla que tengamos fuentes compactas, mucho menores que la resolución del beam, enmascaradas en un ruido homogéneo que domina sobre estas fuentes. El uso más probable de este código es para la detección de fuentes compactas en imágenes del FCM realizadas por misiones futuras, predecesoras a Planck como puede ser la colaboración QUIJOTE.

Para finalizar queda comentar que el código implementado ha sido alojado en un repositorio GitHub. El repositorio está a nombre del estudiante pero reconoce las autorías del director del trabajo. En el se encuentran los catálogos obtenidos para las frecuencias del LFI, todo el src con el resto de programas necesarios para correr el método, y un fichero texto README en el que se explica el uso y el tipo de datos que necesita como input y que genera como output.

Repositorio: <https://github.com/Chemap98/Compact-sources-detection-on-CMB-images>.

A Misión Planck

La misión Planck ha supuesto un gran avance en la astronomía de microondas y el estudio de CMB. El trabajo que hemos realizado se ha inspirado en varios de los trabajos de la misión Planck. Por todo esta vamos a dedicar este apéndice a mencionar aspectos importante de esta misión como cronología, características, objetivos, e hitos.

Planck ha medido las variaciones de temperatura del FCM con una sensibilidad, resolución angular y rango de frecuencias sin precedentes. Ha aportado las imágenes más precisas hasta la fecha del Universo joven, cuando tenía solo 380.000 años de edad.

El satélite medía 4,2 metros de altura y tenía un diámetro máximo de 4,2 metros. Pesaba aproximadamente 1900 kg. Fue puesto en órbita el 14 de mayo del año 2009. Desde entonces y hasta el 23 de octubre cuando se desconectó Planck realizó siete barridos de cielo (hasta agotar el helio que lo refrigeraba). Orbitó la órbita de Lissajous en el segundo punto de Lagrange (L2). Esta misión fue la primera misión europea en estudiar el FCM.

Planck contaba con dos experimentos: el HFI (Instrumento de alta frecuencia) que realizó cinco barridos a frecuencias entre 100 y 857 GHz hasta quedarse sin helio en enero del 2012. Por otra parte el LFI (Instrumento de baja frecuencia) realizó durante toda la duración del experimento siete barridos a las frecuencias de 30, 44, y 70 GHz.

Los objetivos de Planck fueron los de mapear el FCM con mejoras en la sensibilidad y resolución angular, y probar teorías sobre el nacimiento y evolución del Universo.

Planck ha tenido que detectar y analizar las emisiones de fondo que se interponen entre el FCM y nosotros. Los primeros datos fueron publicados en enero de 2011, y los primeros resultados cosmológicos se publicaron en marzo de 2013.

Planck a proporcionado mediciones con una precisión marcada por límites fundamentales de la astrofísica, por ello es imposible tomar mejores imágenes que las tomadas por Planck. La alta sensibilidad de Planck ha permitido tomar las mejores imágenes de las anisotropías del FCM hasta la fecha, permitiendo a los científicos aprender más sobre la evolución del Universo.

Planck es una misión de la ESA, la agencia espacial europea del espacio. Esta agencia está compuesta por 18 miembros. Entre estos 18 países que forman parte de la ESA se encuentra España cuyos científicos, entre ellos el director de este trabajo Diego Herranz y compañeros del Instituto de Física de CAntabria (IFCA), han aportado grandes contribuciones al desarrollo de la misión.

En 2015 y 2018 Planck presentó nuevos resultados. Además de los mapas en microondas de la temperatura del FCM Planck ha medido también su polarización. Entre las publicaciones de Planck⁵ se ha dado información sobre la calibración de los experimentos, separación de componentes, espectro de potencia del FCM, parámetros cosmológicos, inflación, o efectos de lente gravitacional entre otros.

En conclusión: Planck ha aportado los datos más precisos sobre el Universo temprano, mejorando a sus antecesores como WMAP o COBE. Es una macro colaboración de científicos europeos que han permitido mejorar enormemente nuestros conocimientos del cielo en las microondas así como la cosmología.

Referencias

- [1] E. L. Wright, “WMAP First Year Results”, *arXiv e-prints*, astro-ph/0306132, astro-ph/0306132, Jun. 2003. arXiv: astro-ph/0306132 [astro-ph].
- [2] Planck Collaboration, P. A. R. Ade, N. Aghanim, *et al.*, “Planck early results. I. The Planck mission”, *Astronomy and Astrophysics*, vol. 536, A1, A1, Dec. 2011. DOI: 10.1051/0004-6361/201116464. arXiv: 1101.2022 [astro-ph.IM].
- [3] W. Hu and S. Dodelson, “Cosmic Microwave Background Anisotropies”, *Annual Review of Astronomy and Astrophysics*, vol. 40, pp. 171–216, Jan. 2002. DOI: 10.1146/annurev.astro.40.060401.093926. arXiv: astro-ph/0110414 [astro-ph].
- [4] J. Bolton, G. Stanley, and O. Slee, “Positions of three discrete sources of galactic radio-frequency radiation”, *Nature*, vol. 164, no. 4159, pp. 101–102, 1949.
- [5] A. A. Penzias and R. W. Wilson, “A measurement of excess antenna temperature at 4080 mc/s.”, *The Astrophysical Journal*, vol. 142, pp. 419–421, 1965.

⁵<https://www.cosmos.esa.int/web/planck/publications>.

- [6] G. de Zotti, M. Massardi, M. Negrello, and J. Wall, “Radio and millimeter continuum surveys and their astrophysical implications”, *The Astronomy and Astrophysics Review*, vol. 18, no. 1-2, pp. 1–65, Feb. 2010. DOI: 10.1007/s00159-009-0026-0. arXiv: 0908.1896 [astro-ph.CO].
- [7] S. M. Leach, J. -.-F. Cardoso, C. Baccigalupi, *et al.*, “Component separation methods for the PLANCK mission”, *Astronomy and Astrophysics*, vol. 491, no. 2, pp. 597–615, Nov. 2008. DOI: 10.1051/0004-6361:200810116. arXiv: 0805.0269 [astro-ph].
- [8] J. Högbom, “Aperture synthesis with a non-regular distribution of interferometer baselines”, *Astronomy and Astrophysics Supplement Series*, vol. 15, p. 417, 1974.
- [9] E. Bertin and S. Arnouts, “SExtractor: Software for source extraction”, *Astronomy and Astrophysics Supplement Series*, vol. 117, no. 2, pp. 393–404, 1996.
- [10] D. Herranz and P. Vielva, “Cosmic microwave background images”, *IEEE Signal Processing Magazine*, vol. 27, no. 1, pp. 67–75, Jan. 2010. DOI: 10.1109/MSP.2009.934716. arXiv: 1101.0707 [astro-ph.CO].
- [11] G. De Zotti, M. Bonato, M. Negrello, *et al.*, “Extragalactic astrophysics with next-generation CMB experiments”, *Frontiers in Astronomy and Space Sciences*, vol. 6, 53, p. 53, Aug. 2019. DOI: 10.3389/fspas.2019.00053. arXiv: 1907.05323 [astro-ph.GA].
- [12] M. López-Caniego, “The QUIJOTE experiment”, in *IAU General Assembly*, vol. 29, Aug. 2015, p. 2257869.
- [13] D. J. Fixsen, “The temperature of the cosmic microwave background”, *The Astrophysical Journal*, vol. 707, no. 2, pp. 916–920, Nov. 2009, ISSN: 1538-4357. DOI: 10.1088/0004-637x/707/2/916. [Online]. Available: <http://dx.doi.org/10.1088/0004-637x/707/2/916>.
- [14] K. M. Górski, E. Hivon, A. J. Banday, *et al.*, “HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere”, vol. 622, no. 2, pp. 759–771, Apr. 2005. DOI: 10.1086/427976. arXiv: astro-ph/0409513 [astro-ph].
- [15] A. Zonca, L. Singer, D. Lenz, *et al.*, “Healpy: Equal area pixelization and spherical harmonics transforms for data on the sphere in python”, *Journal of Open Source Software*, vol. 4, no. 35, p. 1298, Mar. 2019. DOI: 10.21105/joss.01298. [Online]. Available: <https://doi.org/10.21105/joss.01298>.
- [16] J. C. Bancroft, “Introduction to matched filters”, *CREWES Research*, vol. 297, 2002.
- [17] J. G. Proakis and D. G. Manolakis, “The discrete fourier transform: Its properties and application”, *Digital Signal Processing*, pp. 464–472, 2009.
- [18] J. Delabrouille, M. Betoule, J. -.-B. Melin, *et al.*, “The pre-launch Planck Sky Model: a model of sky emission at submillimetre to centimetre wavelengths”, *Astronomy and Astrophysics*, vol. 553, A96, A96, May 2013. DOI: 10.1051/0004-6361/201220019. arXiv: 1207.3675 [astro-ph.CO].
- [19] M. López-Caniego, D. Herranz, J. González-Nuevo, *et al.*, “Comparison of filters for the detection of point sources in Planck simulations”, vol. 370, no. 4, pp. 2047–2063, Aug. 2006. DOI: 10.1111/j.1365-2966.2006.10639.x. arXiv: astro-ph/0606199 [astro-ph].
- [20] Planck Collaboration, N. Aghanim, C. Armitage-Caplan, *et al.*, “Planck 2013 results. IV. Low Frequency Instrument beams and window functions”, *Astronomy and Astrophysics*, vol. 571, A4, A4, Nov. 2014. DOI: 10.1051/0004-6361/201321544. arXiv: 1303.5065 [astro-ph.CO].
- [21] Planck Collaboration, P. A. R. Ade, N. Aghanim, *et al.*, “Planck 2015 results. XXVI. The Second Planck Catalogue of Compact Sources”, *Astronomy and Astrophysics*, vol. 594, A26, A26, Sep. 2016. DOI: 10.1051/0004-6361/201526914. arXiv: 1507.02058 [astro-ph.CO].
- [22] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography”, *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, ISSN: 0001-0782. DOI: 10.1145/358669.358692. [Online]. Available: <https://doi.org/10.1145/358669.358692>.
- [23] J. E. R. P. T. Boggs, “Orthogonal distance regression, Statistical analysis of measurement error models and applications: Proceedings of the ams-ims-siam joint summer research conference held june 10-16, 1989”, *Contemporary Mathematics*, vol. 112, no. 1, p. 186, 1990.
- [24] G. de Zotti, R. Ricci, D. Mesa, *et al.*, “Predictions for high-frequency radio surveys of extragalactic sources”, *Astronomy and Astrophysics*, vol. 431, no. 3, pp. 893–903, Mar. 2005. DOI: 10.1051/0004-6361:20042108. arXiv: astro-ph/0410709 [astro-ph].

- [25] M. Tucci, L. Toffolatti, G. de Zotti, and E. Martínez-González, “High-frequency predictions for number counts and spectral properties of extragalactic radio sources. New evidence of a break at mm wavelengths in spectra of bright blazar sources”, *Astronomy and Astrophysics*, vol. 533, A57, A57, Sep. 2011. DOI: 10.1051/0004-6361/201116972. arXiv: 1103.5707 [astro-ph.CO].
- [26] G. de Zotti, M. Massardi, M. Negrello, and J. Wall, “Radio and millimeter continuum surveys and their astrophysical implications”, *The Astronomy and Astrophysics Review*, vol. 18, no. 1-2, pp. 1–65, Feb. 2010. DOI: 10.1007/s00159-009-0026-0. arXiv: 0908.1896 [astro-ph.CO].
- [27] Planck Collaboration, Y. Akrami, F. Argüeso, *et al.*, “Planck intermediate results. LIV. The Planck multi-frequency catalogue of non-thermal sources”, *Astronomy and Astrophysics*, vol. 619, A94, A94, Nov. 2018. DOI: 10.1051/0004-6361/201832888. arXiv: 1802.08649 [astro-ph.CO].